UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PUBLIC POLICY APPLICATIONS OF REGRESSION DISCONTINUITY DESIGN

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

DOCTOR OF PHILOSOPHY

in

ECONOMICS

by

Eva S. Shapiro

June 2019

The Dissertation of Eva S. Shapiro
is approved:

Professor Carlos Dobkin, Chair

Professor Robert Fairlie

Professor Justin Marion

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Public Policy Applications of Regression Discontinuity Design

by

Eva S. Shapiro

This thesis consists of three projects based on the application of Regression Discontinuity Design (RDD) analysis to questions of public economics. Chapter 1 evaluates the effects of English Language Learner programs on student achievement. There is a significant and well-documented academic achievement gap between native English speakers and English language learners (ELLs) in US public schools. [1] This gap is particularly large in California, the US state with the largest population of ELLs. As a result of being classified as having limited English proficiency, ELLs follow a different track through the primary and secondary public school system in the U.S., according to federal requirements and implemented at the state level. The effect of this differential treatment of non-native English speakers is ambiguous, and the division of students into separate tracks based on English lacks support in existing literature. I use longitudinal microdata from a large northern California school district to estimate the effects of ELL programs on student outcomes. Students in California are categorized as ELLs based on performance on a standardized assessment of English language proficiency, the California English Language Development Test (CELDT). Students classified as are eligible for enrollment in a bilingual programs. For primary school students, this is the Alternative Language Acquisition (ALA) program. I exploit discontinuity in the

probability of enrollment in ALA at the threshold CELDT score for English proficiency. I find that schools do not observe the rule for assignment to ALA based on CELDT score.

Chapter 2 evaluates whether consumers respond to information about product quality in the context of automobile safety. Consumer choice over new vehicles is a function of multiple vehicle attributes, including price, fuel efficiency and safety. However, because vehicle safety is often correlated with other characteristics of vehicle quality, estimating consumer preferences for safety over other attributes is empirically difficult. Using a federal program in the United States that provides public safety ratings for new passenger vehicles, I exploit discontinuity in the assignment of 5-star vehicle safety ratings in continuous probability of injury measurements calculated based on crash test performance. I evaluate whether new vehicle models that just miss a star threshold on the National Highway Transportation and Safety Administration's New Car Assessment Program's 5-Star Rating scale experience lower national sales volumes relative to vhicles that just exceed the ratings threshold.

Chapter 3 evaluates the accuracy of the New Car Assessment Program (NCAP) safety ratings in predicting real-world crash outcomes. Regulatory policies such as safety standards for seat belts or airbags are aimed at improving vehicle safety and reducing occupant injury and fatality rates, but may distort driver and occupant behavior. Vehicle safety ratings can provide standardized, transparent and comparative information to buyers. However, the value of any safety rating to consumers depends on the accuracy of the rating regime. In the context of transportation policy, this means real-world loss of

life, injury and property damage on a national scale. In the U.S., the National Highway and Transportation Administration (NHTSA) evaluates the safety of all new vehicles sold in the United States and publishes these safety ratings via the New Car Assessment Program (NCAP), a program which is emulated internationally. Using a novel dataset with the continuous underlying running variable, probability of injury, used to calculate the 5-Star NCAP safety rating seen by consumers, I evaluate whether U.S. NCAP safety ratings accurately predict real-world crash outcomes in terms of vehicular damage, personal injury and loss of life. Matching NCAP rating with crash report data from the U.S. Fatality Analysis Reporting System and Texas' Crash Records Information System, I find minimal correlation between NCAP rating and real-world crash outcomes.

In summary, this dissertation applies Regression Discontinuity Design analysis to public education policy in the context of the CELDT and to regulatory and transportation policy in the context of NCAP. Successful implementation of these policy regimes depends on accuracy of classification, whether of students or vehicles. Failures of evaluation and enforcement result in distortionary effects with real-world effects on educational attainment and motor vehicle safety. Regression Discontinuity Design provides a powerful tool for evaluating the true efficacy of public policies.

To my grandmother,

Alma Boyer.

Acknowledgments

This dissertation would not have been possible without the patience and support of my committee members, Robert Fairlie and Justin Marion, and in particular that of my advisor, Carlos Dobkin. Thank you for supporting me as not only a fledgling economist but also a human. I would also like to thank Rodney Ogawa for introducing me to and facilitating my work with education practitioners. This dissertation would not have been possible without the contributions of Jeremy West and James Pettit. Special thanks to Sandra Reebie in the UCSC Economics Department. Finally, boundless appreciation to my family for a lifetime of love and support.

Chapter 1

The California English Language Development Test, English Language Learner Programs and Student Achievement

1.1 Introduction

The number of students in U.S. public schools who are non-native English speakers is substantial and increasing. Nearly one quarter of the 35 million students in U.S. public elementary schools and the 15 million students in U.S. public high schools are Hispanic, and the majority of these students speak Spanish as their home language. [1] Demographic trends such as birth and fertility rates among white and Hispanic populations suggests that the number and fraction of non-native English speakers in

the nations public schools will only increase over the next decade. [5]

Students for whom English is not a native language face acute obstacles in the U.S. public education system that differentiate them from their English-speaking peers. Their ability to benefit from schooling in English-speaking classrooms is compromised by their lack of English language proficiency, often putting non-English speakers at a disadvantage relative to native English speakers. There is a significant and well-documented achievement gap between native English speakers and native Spanish speakers. On both the reading and writing components of the NAEP, there is a significant achievement gap between Whites and Hispanics across all states. [1] Students of Spanish-speaking backgrounds have higher dropout rates and are more likely to be placed in lower ability classes and special education programs. [36]

These trends are reflected in long-term outcomes such as college enrollment, career outcomes, and earnings, contributing to growing education and income inequality between whites and Hispanics in the U.S. [32] Hispanics earn significantly lower wages than whites, particularly for individuals with low levels of education levels. [**?**] Differences in language skill decrease substitutability between white and Hispanic workers, further contributing to the wage gap between the two populations. [53]

This language-based gap in ability to learn raises questions of equal treatment within the public education system. In an effort to recognize and aid these students, national regulations require that students be tested in English language proficiency and receive specialized assistance if they are identified as falling below the acceptable threshold for English proficiency. These students are categorized as Limited English

Proficient (LEP). 9% of U.S. primary and secondary school students in the 2010-2011 academic year were classified as LEP, and subsequently enrolled in language assistance programs. [32] National funding for these English language learner (ELL) programs was over $732 million in the 2012 fiscal year. [5]

California serves the most and has the highest percentage of English language learners of any state. [1] 50% of CA public school students are Hispanic, more than double the national average. The majority of these students speak Spanish at home, and native Spanish speakers comprise the vast majority of all ELL students. [5] The achievement gap in California between white and Hispanic students in both reading and mathematics is significantly above the national average. [1] Hispanic students are half as likely as white students to perform at a proficient level on the math section of the California Standards Test (CST), while LEP students are a third as likely. Hispanic students are two thirds as likely as white students to perform at a proficient level on the English language section of the CST, while LEP students are a fifth as likely. This achievement gap is not limited to standardized test scores. English language learners are significantly more likely to be identified as having learning disabilities and placed in special education programs. [77] These academic and program differences have long-term consequences on student development and achievement. 70% of Hispanic students, and 60% of LEP students, graduate from high school in California, relative to 85% of white students.

Unsurprisingly, idenfitying and educating English learners is a central and contentious issue in California. Funding for English learner acquisition in California was

over $160 million in 2012. [32] However, program choice is left largely up to individual counties and districts, and the efficacy of programs targeted at ELLs is ambiguous. A student is classified as an English Language Learner (ELL) based on his or her score on the California Language Development Test (CELDT). The test is administered to non-native English speaking students upon enrollment in the school district. Students who score below the threshold for English proficiency on the CELDT are classified as ELLs. These students are more likely to be placed in a bilingual classroom among peers with significantly lower academic achievement. Classification of a student as an ELL has implications for school assessment criteria and funding. However, my study does not consider the effect of policy and funding incentives on school behavior. I focus on the effect of classification as an ELL on a student's probability of enrollment in a bilingual program.

California Proposition 227, passed in 1998, established English Only (EO) as the primary method of instruction for ELLs. However, programs available to ELLs continue to be designed and implemented at the level of the individual school district. In the school district that I consider in this study, students classified as ELL are eligible for enrollment in a specialized bilingual program, Alternative Language Acquisition (ALA). ALA provides a separate classroom for instruction in Spanish by a certified bilingual instructor. The relative merits of EO and ALA for student development are heavily debated.

Proponents of separate ELL classrooms in which students are instructed in their native language argue students learn best in their native language. In contrast,

critics of bilingual education programs argue that students placed in non-English classrooms are segregated from their English-speaking peers and may never develop the English language skills necessary to succeed in mainstream English classrooms in public schools. Given this disparity in support for ELLs in California and across the U.S., the question of how education policy be structured to best serve ELLs and decrease the long-term and systematic achievement and wage gap between whites and Hispanics seems central to future education policy. Education is a critical component of human capital development, suggesting that placement in ELL programs early in life may have long-term effects on individual achievement, career outcomes, and wealth. If tracking students into native English-speaking and non-native English speaking cohorts contributes to and exacerbates the achievement gap between the white and Hispanic students, the efficacy of bilingual education programs in California, and across the U.S., may need to be reevaluated.

Despite the attention that bilingual education has received in the media and political discussions, relatively little analysis has been done to determine the effect of programs such as ALA. Current research in this area is limited due to self selection and the lack of control groups for bilingual programs. Most estimates of program efficacy are limited by the fact that students are placed in ELL programs based on English proficiency, which is correlated with other observable and unobservable characteristics such as innate and cognitive ability, language acquisition skills, parental support and home environment. This can downwardly bias estimates of the effect of ELL programs on long-term academic outcomes. As a result, isolating the true causal effect of bilingual

programs from the contribution of individual student characteristics is often impossible. This difficulty in estimating the causal effect of bilingual programs on student achievement is compounded by the lack of detailed longitudinal data on student demographics, program enrollment, and academic outcomes.

I overcome these limitations by using a new, longitudinal, student-level data set from an undisclosed large school district in California. In this district, students classified as ELLs can be enrolled in either an English Only (EO) classroom in which they are instructed entirely in English alongside native English speakers, or are enrolled in an Academic Language Acquisition (ALA) program in which they are instructed in Spanish by teachers with bilingual certification in a separate classroom. Eligibility for ALA is determined by performance on the California Language Development Test (CELDT).

In order to evaluate the causal effect of ALA on long-term student outcomes, I use a fuzzy regression discontinuity design (RDD). The fuzzy RDD method exploits discontinuity in probability of enrollment in ALA at the threshold score for English proficiency on the CELDT. Students scoring just above and below the proficiency threshold are likely to exhibit similar observed and unobserved characteristics that are correlated with academic achievement. As a result, whether a student scores above or below the CELDT threshold can be used as an instrumental variable for ALA enrollment in order to estimate the causal effect of ALA on academic outcomes such as course grades and standardized test scores.

I find that students enrolled in ALA score significantly lower on both CST

English Language Arts and Mathematics standardized test scores. The CST is an annual test administered to all students in grades 2 through 11. I report results for CST grades 2-8 scores. ALA students also achieve significantly lower grades in middle school. However, schools in this district appear to systematically violate the rule for assignment to ALA. Less than 50% of students who score below the threshold for English proficiency on the CELDT are enrolled in ALA. This is robust to conditioning on ALA availability by school and to demographic and age subgroups. This level of fuzziness in my RD design restricts my ability to draw conclusions about the causal effect of ALA vs. EO programs on students achievement.

The remainder of this Chapter is structured as followed. Section 2 provides a brief review of the literature on language development programs and student achievement. Section 3 discusses the institutional context and background of ELL programs in California and the CELDT. Section 4 describes the data used and provides summary statistics. Section 5 introduces the model and methodology that I use to isolate the effects of English language development programs on student outcomes. Section 6 provides a preliminary analysis of the data. Section 7 concludes. Section 8 discusses future work in this area, and Section 9 provides a timeline for progression.

1.2 Previous Literature

Evaluation of bilingual education programs is compromised by the availability of longitudinal, student-level data and endogeneity of covariates with assignment to

programs. For these reasons, there are a limited number of studies evaluating the causal effects of bilingual education programs on student achievement. However, there is a rich literature on tracking and peer effects.

The effect of tracking students into separate academic programs based on English proficiency is unclear in education economics theory. Tracking high and low performing students separately could be harmful if less qualified teachers are assigned to low performing students, as well as if there are positive peer effects from being in the same class as high-performing students. However, tracking could be beneficial by allowing teachers to focus on heterogeneous student needs and by modifying teacher incentives. This raises the question of whether lower achieving students or students of a non-English background do better when placed among similarly low-achieving or non-English-speaking peers, when a teacher can focus on their needs and design an adequately paced curriculum for their language acquisition, or whether lower achieving or Spanish-speaking students learn more effectively among higher achieving or English-speaking peers. A number of studies have focused on peer effects and tracking students by academic ability.

Duflo et al. (2008) consider a tracking experiment in 210 schools in Western Kenya. [31] Schools were given funding to offer a second class section, and students were allocated to class sections either randomly or based on ability. The authors find significant, positive, and persistent effects of tracking in both the high-performing and low-performing class sections. They argue that these effects are due largely to the behavioral responses of teachers. The authors find no support for the hypothesis that

tracking might harm lower ability students by removing them from classrooms with higher ability peers.

A number of studies in the U.S. have focused specifically on the effect of tracking students with particular groups of their peers. Burke & Sass (2013) analyze classroom peer effects on student achievement. [17] Using longitudinal data for students in grades 3-10 in Florida public schools, they find evidence of peer effects using non-linear models. Their results suggest that there is an optimal allocation of students to classrooms that would maximize individual student achievement.

Carrell et al. (2011) analyze peer effects using students at the U.S. Air Force Academy (USAFA). [21] The authors assign half of entering freshmen in 2007 and 2008 at the USAFA into peer groups optimally designed to maximize the grades of the students predicted to be in the lowest third of incoming students in terms of academic achievement. The remaining half of students are randomly assigned to squadrons. However, the authors find that students in the treatment group saw a statistically significant reduction in grades if they were in the lowest third, a statistically significant increase in grades if they were in the middle third, and no effect if they were in the highest third. The authors argue that this is due to sub-group sorting dynamics that prevent the treatment cohorts from functioning as predicted to benefit the lowest achieving students. This scenario is unlikely in kindergarten and elementary school classrooms where the ability of students to identify and select playmates on ability, at least within the classroom setting in the absence of parents, is minimal, and in language development settings, in which peer choice is likely dependent on language proficiency.

Fewer studies have dealt with tracking in the context of language development. Callahan (2005) uses data from a rural high school in northern California to evaluate the effect of track placement on ELLs' GPA, standardized test scores, and performance on the California High School Exit Exam (CAHSEE). [18] Track is defined according to the proportion of classes on a student's transcript meeting college entry requirements. However, students are not randomly assigned to different tracks, and her analysis suffers from a small sample size (n = 355). Cho (2012) finds evidence that ELLs have negative peer effects on reading achievement as measured by test scores. [24] This raises questions of equity and fairness; if all students benefit from being in a classroom with the highest achieving students, are these high-achieving students hurt by being put in a classroom with lower-achieving peers? What does this suggest for ELL programs? Should the least proficient English speakers be put in classrooms with similar students who are learning at the same pace, or which native English speakers with whom they can interact to improve their language ability?

Although English language development is debated heavily in educational, linguistic, and child development literature, there is limited research on the effects of language development and bilingual education programs using econometric methods and student-level data. The California Department of Education has published several studies on language development programs. However, these reports have been primarily program evaluations and cost benefit studies, rather than econometric analyses. Mitchell et al. (1997), in a report for the California Educational Research Cooperative on the Santa Ana Unified School District (SAUSD), review special programs for language de-

velopment. [59] They find that students enrolled in specialized language programs are more likely to achieve English fluency than students enrolled in mainstream classrooms, and that students make more rapid progress transitioning between proficiency levels at low levels than at high levels. However, this report is mainly summary statistics and takes no steps to address endogeneity of assignment to special programs for language development.

Parrish et al. (2006), in a report to the California Department of Education on the effects of Proposition 227, compare English-only students to ELLs and reclassified ELLs using achievement on the Stanford Achievement Test, Version 9 (SAT9) and California Standards Test (CST). [61] Using average student test scores for all students in California from 1998-2001 by language classification, they find that the greatest achievement for ELLs was in bilingual programs. However, they use aggregated rather than student-level data over a short assessment period, and are unable to include demographic controls.

Several studies focus on the resource and achievment gap between ELL and English-speaking students. Gandara et al. (2003) argue that ELLs receive unequal education services such as assignment to less qualified teachers, inferior facilities and curriculum, segregation from English-speaking peers, and assessment via invalid instruments. [35] Using the Early Childhood Longitudinal Study (ECLS) of the U.S. Department of Education, the American Institutes for Research Implementation of Proposition 227 Study, and data from the California Department of Education, the authors argue that the academic achievement of ELLs falls considerably behind that of native English

speakers. They show a significant achievement gap between English learners and English speakers on the SAT9 between 1998 and 2002. Using the ECLS data, the authors argue that this is partly attributable to the fact that ELLs begin kindergarten significantly behind English-speakers in terms of language development, mathematics and general knowledge assessments. Further, the authors find that ELLs are more likely to be placed in special education programs. If this is correct, then ELL students are at a significant disadvantage relative to their English-speaking peers, and estimates of the effects of ELL programs on academic outcomes will be downwardly biased.

A number of studies look specifically at the effect of two-way bilingual programs on academic outcomes. Shneyderman and Abella (2009) study the effects of a TWBI program on Spanish language proficiency and reading and mathematics achievement in English in a school district in the Southeastern U.S. [75] Using a four academic-year period, the authors find that students enrolled in the TWBI program had reading and mathematics achievement as high as or higher than similar students not enrolled in the program. Although the authors construct a synthetic comparison group via a matching technique, their results are compromised by selection bias and endogeneity, upwardly biasing their estimates of the effect of TWBI programs on student achievement.

Alanis (2000) studies the effects of a TWBI program in Texas on linguistic and academic achievement. [9] Her sample incudes fifth-grade students of Mexican origin who were either native Spanish or native English speakers and were enrolled in the TWBI program for a minimum of three years. She finds that TWBI students performed at the same level or at a higher level than students who did not participate in the TWBI

program on the Texas Assessment of Academic Skills (TAAS). However, her sample size is only 85 and, similarly to Shneyderman and Abella, because students are not randomly assigned to the program, her study is compromised by issued of selection bias that upwardly bias her estimates of the program effects.

Most relevant to my analysis are studies that rely on natural experiments and regression discontinuity design (RDD) settings to evaluate the effects of ELL programs. Barnett et al. (2007) compare the effects of a two-way bilingual immersion (TWBI) and an English-only (EO) preschool program on learning in an undisclosed city in the Northeast. [13] Three- and four-year olds from both native English- and Spanish-speaking backgrounds were randomly assigned to either the an EO program or a TWBI program that alternated weekly between English and Spanish. The authors find significant improvement in Spanish vocabulary among native Spanish speakers in the TWBI relative to the EO program, but otherwise mixed effects. However, the authors are unable to track academic outcomes beyond a year after initial assignment, leaving the long-term effects of TWBI programs unclear.

Further, these TWBI programs are not directly comparable to the ALA programs in the school district I consider. Students in ALA are taught alsmost entirely in Spanish, particularly in younger grades. In this sense, ALA is not a strictly bilingual program. The school district I consider does have a TWBI program, but only a small minority students are enrolled, and the TWBI classroom languages are generally not Spanish, but other languages such as Mandarin and French.

Chin et al. (2012) uses an RDD based on class size to evaluate the effect of

bilingual education programs in Texas. The authors investigates a Texas education law requiring that a school district offer bilingual education when its enrollment of limited English proficient (LEP) students in a given elementary grade and language is at least twenty. [23] Using a RDD around this cutoff, the authors estimate the causal effect of enrollment in a two-way bilingual program on academic achievement by LEP and non-LEP students. They find no effects of bilingual education programs on standardized test scores of native Spanish speaking students, but some evidence of positive spillovers to native English speakers. This could be due to positive peer effects for LEP students from having English speaking peers, further suggesting a benefit of EO over ELL programs.

Robinson (2011), using data from California, also uses an RDD to evaluate the effect of language development programs. However, he focuses on the effect of reclassification of ELL students as Reclassified English Proficient (REP) on student achievement. [67] His sample uses data from an undisclosed district in California. Because reclassification generally involves a change in instruction, including decreased language development support, this can have an adverse effect on student achievement. The author uses a binding score cutoff in reclassification eligibility to evaluate these effects using a sample of students in grades 4 through 10. He finds no effect of reclassification on academic outcomes of ELL students. However, this could be due to positive peer effects associated with being enrolled in an English only classroom, suggesting the non-EO programs are less beneficial to the English language and academic development of ELLs than EO classes. Further, he is unable to analyze the effects of ELL programs on students first enrolling in the district at younger ages, and his results likely suffer from

endogeneity from previous years of schooling in the district. In addition, his sample does not focus on Hispanic students; over one third of the ELL students in his sample are Asian. Finally, Robinson only looks at student outcomes up to one year from the date of reclassification, and is unable to observe long-term outcomes such as CST scores more than a year from the date of reclassification, middle school grades, or PSAT score.

Finally, Matsudaira (2009) also uses a RDD around the ELL reclassification threshold using data from an undisclosed large urban school district in the northeast U.S. [55] As in the California school district I study, students enrolling in the district Matsudaira studies are given a language assessment test if they report on a home survey that the students native language or the language spoken at home is not English. Students scoring below the 40th percentile on this test are classified as Limited English Proficient (LEP). The ultimate decision of whether these students are enrolled in a bilingual education program is then up to the individual students teacher and parents, and subject to availability at the students school. He finds no effect of reclassification on academic outcomes within two years of relcassification testing date. He finds that students scoring below the threshold for English proficiency are 90 percent more likely to be placed in a bilingual program. However, he finds no difference in academic achievement in reading and math between the two groups. Although I do not focus on the effects of reclassification in this paper, future work and data availability makes this a potential extension of my research.

To my knowledge, no studies employing econometric techniques use California data to evaluate the effect of initial placement in ALA on long-term academic outcomes.

[33] Access to a new, comprehensive, student-level, longitudinal dataset from a large California school district provides a unique and novel opportunity to evaluate the effect of bilingual education programs on students' academic achievement.

1.3 Institutional Context

1.3.1 California English Language Development Test

U.S. states receive federal funding for ELL programs under Part A of Title III, the English Language Acquisition, Language Enhancement, and Academic Achievement Act. Under the 2001 No Child Left Behind policy, schools receiving Title III federal funds are required to conduct annual tests of reading and mathematics for all students. Schools may exempt English language learners (ELLs) from achievement testing in English for up to three years, but are required to assess English language proficiency annually and without an exemption period. [36] In California, this assessment of ELLs is done via the California English Language Development Test (CELDT). The CELDT was developed by CTB-McGraw Hill to fulfill California Assembly Bill 748 (1997) and California Senate Bill 638 (1999), which require schools to assess the English language development of English learners, and has been administered to ELLs in California since 2001. According to California Education Code, Section 60810(d), the purposes of the CELDT are: (1) To identify pupils who are limited English proficient. (2) To determine the level of English language proficiency of pupils who are limited English proficient. (3) To assess the progress of limited-English-proficient pupils in acquiring the skills of

listening, reading, speaking, and writing in English.

The CELDT is administered to all students in California whose native language is not English. The exam is administered either as an initial assessment (IA) to students newly enrolled in the district, as determined by a home language survey given to parents, or as an annual assessment (AA) to students classified as Limited English Proficient (LEP) based on a previous year's CELDT score. As a result, many students have CELDT scores for multiple years, beginning with their year of enrollment in the district.

The CELDT exam has four subtests: listening, speaking, reading, and writing. Grades K-1 are assessed only in listening and speaking, while grades 2-12 are assessed in all categories. Prior to 2006, the listening and speaking subsections were combined into a single Listening & Speaking (L & S) section. These subtests contained questions in the following areas:

1. Listening and Speaking: following oral directions, phonemic awareness and control, oral vocabulary, and story retelling. These were in the form of multiple choice and dichotomous constructed response.

2. Reading: word analysis, fluency and vocabulary, comprehension and analysis. These were in the form of multiple choice.

3. Writing: grammar and structure, writing sentences, and short comprehension. These were in the form of multiple choice, short answer constructed response, and extending writing constructed response.

Students are assessed along a continuum of five levels: Beginning (1), Early

Intermediate (2) Intermediate (3), Early Advanced (4) and Advanced (5).

1. Advanced: Students performing at this level of English language proficiency communicate effectively with various audiences on a wide range of familiar and new topics to meet social and academic demands. In order to attain the English proficiency level of their native English-speaking peers, further linguistic enhancement and refinement are necessary.

2. Early Advanced: Students performing at this level of English language proficiency begin to combine the elements of the English language in complex, cognitively demanding situations and are able to use English as a means for learning in other academic areas.

3. Intermediate: Students performing at this level of English language proficiency begin to tailor the English language skills they have been taught to meet their immediate communication and learning needs.

4. Early Intermediate: Students performing at this level of English language proficiency start to respond with increasing ease to more varied communication tasks.

5. Beginning: Students performing at this level of English language proficiency may demonstrate little or no receptive or productive English skills. They may be able to respond to some communication tasks.

For a student to be classified as English proficient, he or she must achieve an overall CELDT score of at least Level 4 (Early Advanced) and score at least Level

3 (intermediate) on each of the individual subtests: listening, speaking, reading, and writing (or, for Kindergarten and first grade students, just listening and speaking).

1.3.2 English Language Learner Programs

At the elementary school level, students who score Beginning, Early Intermediate or Intermediate on their initial assessment (IA) of the CELDT are eligible for enrollment in a bilingual education program, Alternative Language Acquisition (ALA). ALA classes are taught entirely in Spanish and are only at the elementary school level. ELL students who are not enrolled in ALA receive English-Only instruction, also known as Structured English Immersion (SEI). In theory, EO courses are taught nearly entirely in English, with the curriculum and presentation designed for students who are learning English. [3] However, EO instruction is done within the context of a mainstream English classroom, in which ELLs are taught alongside native English speakers. The exact level of instruction customized to students learning English varies by school, classroom, and teacher, and is not observable in my data.

Proposition 227 of 1998 established EO as the dominant instructional approach for English language development in California. However, at the time ELL students enroll in the district and receive their initial CELDT level, their parents are provided with information about ALA and EO and are able to choose the option they want for their child. All 27 elementary schools in the district have EO programs, but only thirteen have both EO and ALA. As a result, under the Elementary Transfers for ALA Bilingual Program, parents whose children are initially enrolled in an EO program or

whose assigned school does not have an ALA program can apply for an Exception Waiver for a transfer to an ALA classroom or school with available space. If no space is available, the student is placed in an EO classroom. Although parents can choose to have their child enter a mainstream English classroom, most parents who are new to the district and themselves English learners choose to keep their child in an ELL program in which they are able to communicate with the teacher themselves in their native language. Nonetheless, the role of parent choice introduces significant selection bias.

Students enrolled ELL programs who reach middle school while still being classified as an ELL enter English Language Development (ELD) or, for students scoring at the intermediate level, Specially Designed Academic Instruction in English (SDAIE) programs. At the high school level, SDAIE students are enrolled in a special section for regular English classes. Although ELL programs at the middle and high school level are not the focus of this paper, future work using this data could consider ELD and SDAIE programs. However, given previous exposure to ALA and EO programs, there are empirical challenges to disentangling the effects of ELD and SDAIE programs from earlier language development programs.

1.3.3 Reclassification

Students who initially score below proficient on their initial CELDT exam can become Reclassified English Proficient (REP) if their English proficiency improves. Students who have scored below proficient on the CELDT continue to take the exam

each year as an AA. Once students reach the Early Advanced level, they may qualify for reclassification depending on their CST ELA performance and whether they pass the District's writing performance-based assessment. In order to be reclassified as English proficient (REP), a student must achieve an overall CELDT score at least at the Early Advanced level, and individual subtest scores at least at the Intermediate level for all categories.

In order to be eligible for bilingual reclassification, students must meet the following criteria:

1. Language proficiency at least Limited English;
2. Pass on the most recent Writing Performance Based Assessment (PBA);
3. Most recent CST ELA Scaled Score of 317 or higher;
4. Latest CELDT total level of early advanced or advanced;
5. Latest CELDT listening, speaking, reading and writing subtest levels of intermediate, early advanced, or advanced.

I do not focus on English language proficiency reclassification in this paper. However, future work using this data has the potential to evaluate the effect and timing of reclassification on student achievement.

1.4 Data

I use a new dataset from an undisclosed, large northern California school district, which I will refer to as (NCSD). In California, there are 1,036 school districts and over 10,000 schools. NCSD provides a large, representative sample of the ethnic and language diversity of these districts and schools public schools. NCSD has an annual enrollment of over 30,000 students. Nearly three fourths of NCSD students are from minority backgrounds, and over half of students in the district are Hispanic. NCSD has been historically segregated along racial and language lines; in the mid-1980s, however, a court order led to the de-segregation of the North end of the district, which is of lower income and Hispanic population, and the South end of the district, which is predominantly White and Asian and of higher socioeconomic status. Nearly a quarter of NCSD students are ELLs.

NCSD is comprised of 27 elementary schools, 8 middle schools, and 8 high schools. The NCSD dataset provides detailed information on 112,212 distinct students enrolled in the school district between 1997 and 2012. The dataset, which tracks individual students longitudinally over time as they progress through the district, provides data on enrollment, program participation, demographic characteristics, course information and enrollment, teacher information and credentials, school information, course grades and grade point average (GPA), course schedule, standardized test scores, including PSAT and SAT, CAHSEE scores, graduation status, and college enrollment, among other characteristics.

1.4.1 Summary Statistics & Figures

Table 1.1 shows NCSD enrollment numbers. Enrollment in NCSD is relatively consistent across all years 2004-2010, at approximately 30,000 students. For primary school students, in Kindergarten through Grade 5, enrollment is also relatively consistent, at approximately 13,000 for 2004-2010. Most students are observed over multiple years and grades.

Table 1.1: Enrollment by Year and Grade

	Grade						
	K	1	2	3	4	5	K-12
2004-2005	2,304	1,995	2,107	1,966	2,086	2,306	28,123
2005-2006	2,214	2,004	2,159	1,975	1,976	2,243	27,675
2006-2007	2,274	2,010	2,107	1,991	1,901	2,167	27,288
2007-2008	2,307	2,179	2,163	2,089	1,980	2,205	27,719
2008-2009	2,424	2,129	2,284	2,046	2,175	2,215	28,439
2009-2010	2,349	1,819	2,266	2,133	1,899	2,132	28,312
Total	13,872	12,136	13,086	12,200	12,017	13,268	167,556

Table 1.2 shows the number of students taking the CELDT by academic year and grade. There is some variation in number of students taking the CELDT across years and grades. After the 2006-2007 academic year, the number of students taking the CELDT is highest for Kindergarten students and decreases for higher grade levels. The vast majority (83%) of students taking the CELDT are of low socioeconomic status.

Table 1.2: Students Taking CELDT by Year and Grade

	Grade						
	K	1	2	3	4	5	K-12
2004-2005	377	486	613	700	940	884	10,592
2005-2006	713	726	873	993	863	1,097	10,556
2006-2007	884	968	1,279	995	1,074	908	10,672
2007-2008	1,155	1,249	1,068	1,090	843	758	11,043
2008-2009	1,401	798	1,096	759	684	564	11,124
2009-2010	1,032	1,006	765	610	449	396	9,002
Total	5,562	5,233	5,694	5,147	4,853	4,607	73,603

Table 1.3 gives summary statistics on CELDT scores for all years, 2004-2010, by test section for elementary and Kindergarten students. Scores are normalized to the threshold value for English proficiency for each year, grade level, and subtest. Kindergarten students have significantly lower scores, even allowing for normalization by grade level, across all categories in which they take the test, which are listening and speaking. Prior to 2006, this was given as a single subtest, Listening & Speaking (L&S), while later administrations of the test divided listening and speaking into separate subtests. Kindergarten and first grade students do not take the reading and writing subtests.

	Elementary School Students					Kindergarten Students				
	Obs.	Mean	Std. Dev.	Min	Max	Obs.	Mean	Std. Dev.	Min	Max
Total	11438	-35.22	89.83	-351	224	2271	-43.83	99.59	-326	204
L & S	449	-35.71	93.01	-275	233	52	-69.31	86.39	-238	112
Listening	6480	22.19	91.16	-269	306	1752	7.01	89.48	-189	256
Speaking	6696	31.39	109.71	-319	315	1749	9.74	120.15	-265	315
Reading	4068	-3.63	72.48	-291	209	-	-	-	-	-
Writing	4140	5.58	80.79	-401	223	-	-	-	-	-

Table 1.3: Summary Kindergarten CELDT Scores

Figure 1.1 shows the distribution of CELDT total score for all students taking the CELDT in all grades and years. All scoring types are represented: state initial and state annual, transfer initial and transfer annual, and local initial and local annual. Because students taking the CELDT are native Spanish speakers, it is not surprising that a majority of students taking the CELDT score below the threshold level for English proficiency, and there appears to be some clustering at the bottom of the score distribution.

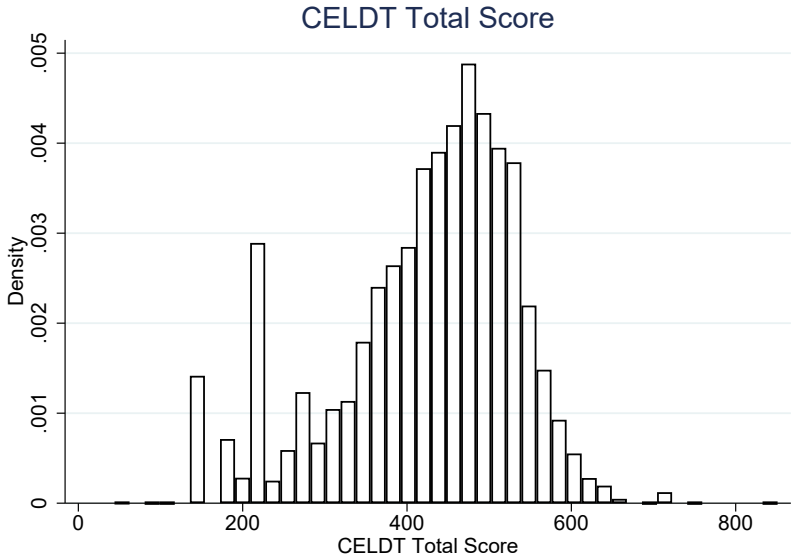


Figure 1.1: Distribution of Total CELDT Score

Students take the CELDT when they are first enrolled in the district, usually in Kindergarten. This yields an initial CELDT score. Students continue to take the CELDT each year until they achieve English proficiency; this yields annual CELDT scores. In addition, CELDT scores are graded by the state Department of Education for official scoring. It is these official, state scores that affect how a student is assessed on

standardized state achievement tests and the funding for ELL programs that a district or school receives. However, it can take several months for the state score to be calculated. As a result, districts provide a local score at the time that the CELDT is administered. It is this local score that determines a student's eligibility for ALA. Thus, the score of greatest relevance for the initial placement of Kindergarten students first enrolling in the district in ALA or EO is the local initial (LI) score. Figure 1.2 shows the distribution of LI CELDT total score for all students taking the CELDT in all grades and years. Because the majority of LI scores are for Kindergarten students, it is not surprising that the vast majority of scores are below the proficiency threshold. Again, there is evidence of clustering at the bottom end of the distribution.

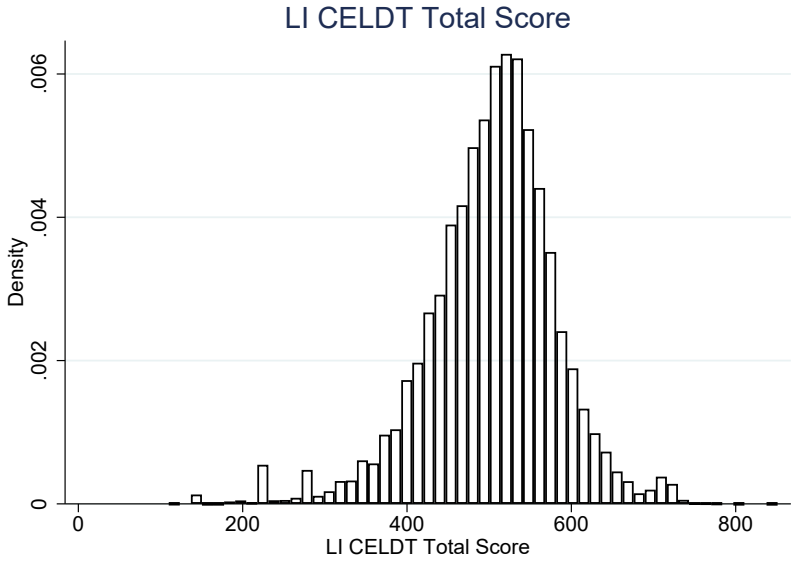


Figure 1.2: Distribution of Local Initial Total CELDT Score

Students receive a total CELDT score, as well as individual scores for each

subtest. In order to be classified as English proficient, students must score above the threshold in each subtest and on the overall CELDT. Figure 1.3 shows the distribution of LI CELDT scores on the Listening & Speaking (L & S) subtest. This subtest was only offered prior to 2006; after 2006, the L & S subtest was divided into separate listening and speaking components. The vast majority of students taking the CELDT prior to 2006, when the L & S subtest was offered, failed this subtest. This may have been one motivation in the reform of the CELDT in 2006.

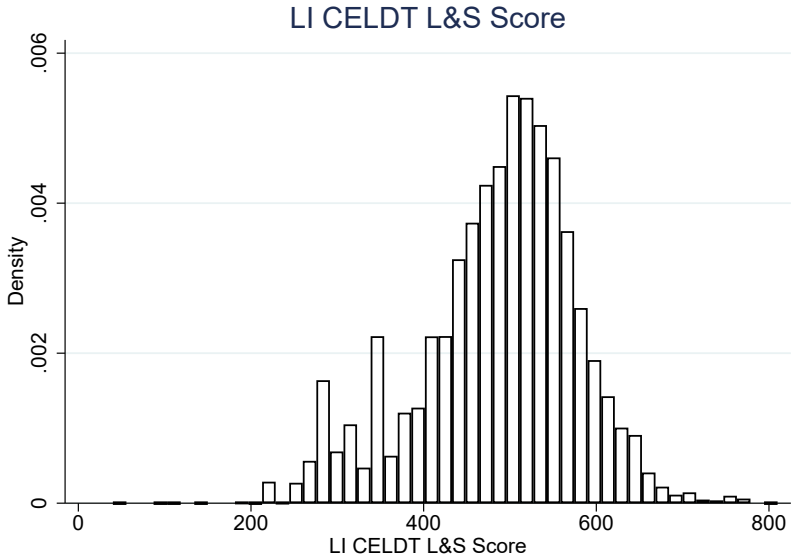


Figure 1.3: Distribution of Local Initial Listening & Speaking CELDT Score

Most students enter the public school system in Kindergarten. Kindergarten students for whom English is not the native language take the CELDT, and receive an initial score. As shown in Figure 1.4, Kindergarten students are particularly likely to fail the CELDT, and there is significant clustering at the bottom end of the score distribution. This is not surprising, given that a significant number of non-English speaking

students enter the public school system with limited English proficiency. However, it also means that most of these non-native English speakers are immediately placed into ALA programs as a result of their CELDT score, an initial placement that has long-term effects on students' trajectories through and performance in higher grade levels.

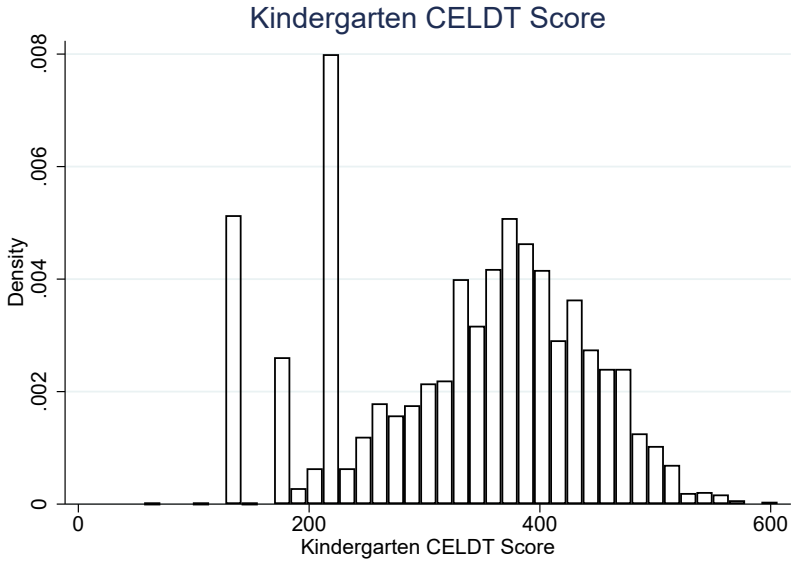


Figure 1.4: Distribution of Kindergarten CELDT Score

Table 1.4 provides summary statistics on demographic information, program participation, CELDT scores, and academic outcomes for students enrolled in English Only (EO) programs and those enrolled in Alternative Language Acquisition (ALA). There are significant differences between EO and ALA students. ALA students are more likely to have taken the CELDT in a slightly later test year, perhaps due to to increased Hispanic and Spanish-speaking populations in the district. Interestingly, ALA students also seem to take the CELDT slightly later in the year, and later in the month, though

the difference is small. ALA students also take the CELDT, on average, at a slightly higher grade. This may suggest that students in ALA are less likely to reclassified as English proficient, and thus more likely to have to continue to take the CELDT at higher grade levels.

	English Only			ALA			Diff: EO - ALA	
	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Diff	p
Test Year	18211	2006.95	2.49	12161	2007.10	2.39	-0.15^{***}	0.00
Test Month	18211	9.94	2.02	12161	10.00	1.81	-0.06^{***}	0.01
Test Day	18211	9.63	10.17	12161	10.29	10.60	-0.66^{***}	0.00
Grade	35052	2.03	1.80	12274	2.08	1.76	-0.05^{***}	0.01
Female	35,052	0.49	0.50	18211	0.50	0.50	-0.01^{**}	0.04
LSES	35052	0.49	0.50	12274	0.92	0.27	-0.42^{***}	0.00
Parent's Ed.	24690	3.34	1.57	8691	2.54	1.99	0.80^{***}	0.00
Hispanic	35,052	0.51	0.50	12274	0.99	0.09	-0.49^{***}	0.00
Spanish	35052	0.30	0.46	12274	0.86	0.35	-0.56^{***}	0.00
GATE	35,052	0.15	0.36	12274	0.07	0.26	0.08^{***}	0.00
Special Ed.	35052	0.08	0.28	12274	0.07	0.26	0.01^{***}	0.00
% ALA	35052	0.24	0.28	12274	0.59	0.11	-0.35^{***}	0.00
Level	18211	3.20	1.20	12161	2.42	1.19	0.78^{***}	0.00
Total Score	5,003	-16.62	77.53	3,194	-74.09	103.42	57.47^{***}	0.00
L & S	214	-22.43	85.32	126	-69.57	106.65	47.14^{***}	0.00
Listening	2,860	35.53	83.12	1,995	4.67	102.22	30.86^{***}	0.00
Speaking	2,955	52.88	94.51	2,010	-4.00	127.20	56.89^{***}	0.00
Reading	1759	6.79	70.25	1,191	-16.63	76.90	23.43^{***}	0.00
Writing	1807	20.78	72.02	1168	-15.71	90.62	36.49^{***}	0.00
CST ELA	35052	435.08	1034.29	12274	410.19	1035.57	24.90^{***}	0.00
CST Math	35052	445.52	986.39	12274	415.00	942.56	30.52^{***}	0.00
MS GPA	13461	2.63	0.89	4,508	2.45	0.79	0.17^{***}	0.00

Table 1.4: Summary Statistics: EO vs. ALA Students

ALA students are more likely to be male, though the difference is small. ALA students are significantly more likely to be of low socioeconomic status (LSES); 92% of ALA students are of LSES, relative to only 49% of EO students. ALA students also have parents with lower levels of education, on average. Unsurprisingly, ALA students are significantly more likely to be Hispanic and speak Spanish as their native language; 99% of ALA students are Hispanic, relative to 51% of EO students, and 86% of ALA

students speak Spanish as their native language, relative to 30% of EO students.

ALA students are half as likely to be enrolled in the Gifted and Talented Education (GATE) program. ALA students are slightly less likely to be enrolled in a Special Education program; however, this may be due to the fact that schools that do not offer ALA programs place students of low English proficiency in special education programs. Students in ALA programs have on average 59% of the students at their school enrolled in ALA, relative to an average of 24% for EO students.

It is clear that students in EO classrooms score significantly higher across all subtests of the CELDT and on total score. EO students score, on average, nearly an entire level higher than ALA students. However, the gap between EO and ALA students is highest in speaking, and lowest in reading. Finally, EO students perform significantly higher on the California Standards Test (CST) in both English Language Arts (ELA) and Math. This suggests that differences between the EO and ALA programs are not limited to language development, but also influence other subjects, such as mathematics. EO students also have higher Middle School GPA, on average.

Not all schools in NCSD offer ALA programs. Figure 1.5 shows the distribution of schools based on proportion of students at the school enrolled in ALA. A number of schools offer no ALA programs. Among schools that do offer ALA programs, most schools have over 50% of students enrolled in ALA. In order to evaluate the effect of ALA vs. EO programs on student achievement, I consider schools that offer ALA programs separately from those that do not.

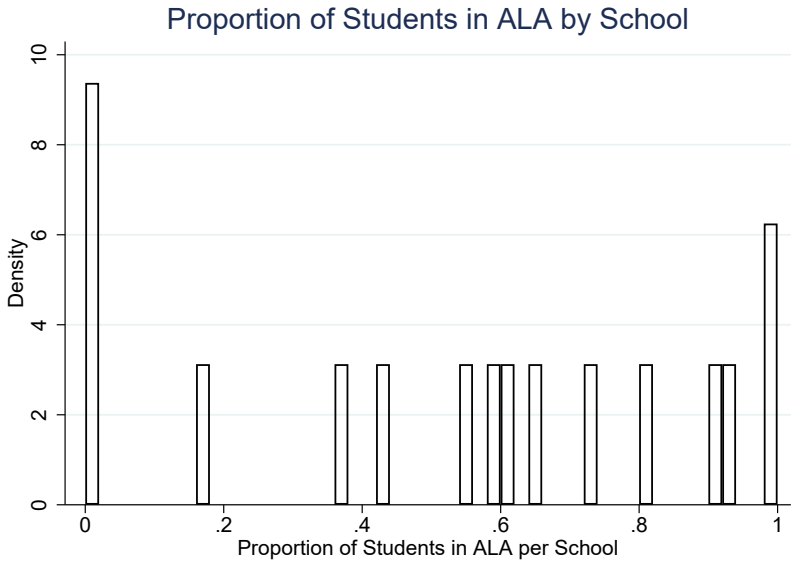


Figure 1.5: ALA Student Share by School

Schools that offer ALA vary significantly from those that do not. Table 1.5 provides summary statistics on demographic information, program participation, CELDT scores, and academic outcomes for schools that offer ALA (high) and those that do not (low). Adjusting the threshold of between low and high above 0 does not significantly change the results. Students at ALA schools take the CELDT in a sightly later aacademic year and later in the month. Students at ALA schools also take the CELDT, on average, at a later grade. Students at ALA schools are slightly more likely to be female.

Students at ALA schools are significantly more likely to be of low socioeconomic status (LSES). 81% of students at ALA schools are LSES, relative to 29% at non-ALA schools. Students at ALA shools are also significantly more likely to be Hispanic and to speak Spanish as their native language. 84% of students at ALA schools

are Hispanic, relative to 32% at non-ALA schools; and 62% of students at ALA schools speak Spanish as their home language, relative to 16% at non-ALA schools.

	Low			High			Diff: Low - High	
	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Diff	p
Test Year	8152	2006.65	2.56	22220	2007.14	2.40	-0.50^{***}	0.00
Test Month	8152	9.98	1.98	22220	9.95	1.92	0.02	0.34
Test Day	8152	8.43	9.55	22220	10.43	10.58	-2.00^{***}	0.00
Grade	18918	1.94	1.78	28408	2.12	1.78	-0.18^{***}	0.00
Female	18918	0.48	0.50	28408	0.49	0.50	-0.01^{***}	0.01
LSES	18918	0.29	0.45	28408	0.81	0.39	-0.52^{***}	0.00
Parent's Ed.	13553	3.77	1.32	19828	2.69	1.82	1.08^{***}	0.00
Hispanic	18918	0.32	0.47	28408	0.84	0.37	-0.52^{***}	0.00
Spanish	18918	0.17	0.37	28408	0.62	0.48	-0.46^{***}	0.00
GATE	18918	0.21	0.41	28408	0.08	0.27	0.13^{***}	0.00
Special Ed.	18918	0.08	0.27	28408	0.08	0.28	0.00	0.16
% ALA	18918	0.00	0.00	28408	0.55	0.13	-0.55^{***}	0.00
Level	8152	3.46	1.17	22220	2.68	1.22	0.78^{***}	0.00
Total Score	2383	-1.02	73.35	5814	-54.59	95.48	53.56^{***}	0.00
L & S	101	-7.14	79.51	239	-53.75	99.62	46.61^{***}	0.00
Listening	1235	45.02	81.79	3620	15.29	94.97	29.73^{***}	0.00
Speaking	1296	61.77	95.05	3669	18.58	115.90	43.19^{***}	0.00
Reading	678	17.76	71.51	2272	-8.76	73.52	26.52^{***}	0.00
Writing	678	36.33	72.70	2297	-2.37	82.25	38.70^{***}	0.00
CST ELA	18918	525.26	1260.18	28408	391.19	881.48	134.07^{***}	0.00
CST Math	18918	526.17	1127.12	28408	428.39	942.97	97.79^{***}	0.01
MS GPA	18918	3.11	0.81	28408	2.49	0.83	0.62^{***}	0.00

Table 1.5: Summary Statistics: High vs. Low ALA-Enrollment Schools

Students at ALA schools are less likely to be enrolled in Gifted and Talented Education (GATE) programs; only 8% of students at ALA schools are enrolled in these programs, relative to 21% at non-ALA schools. There is not a significant difference in proportion of students enrolled in Special Education between High and Low ALA Schools. However, this may be due to greater availability of special education resources at low ALA schools, so does not rule out the possibility that ALA students are more likely to be placed in special education programs.

Students at ALA schools perform significantly worse on all sections of the

CELDT than students at non-ALA schools. This difference is particularly pronounced on the speaking and writing sections. Finally, students at ALA schools have significantly lower academic achievement. They score much worse than students at non-ALA schools on the ELA section of the CST, as well as on the CST mathematics section, though by less. Students at ALA schools also have significantly lower middle school GPA than students at non-ALA schools.

Figure 1.6 shows the distribution of student scores on the English Language Arts (ELA) portion of the California Standards Test (CST). The CST is administered every two years to all students in California, and is the primary assessment by which school and district funding is determined. ALA students perform significantly worse than EO students on ELA CST.

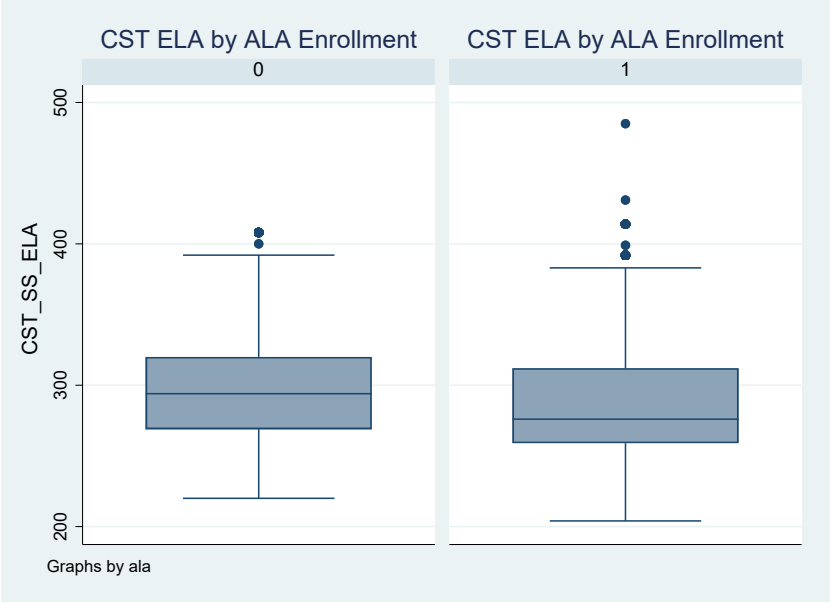


Figure 1.6: CST ELA Performance by ALA Enrollment

The out-performance of ALA students by EO stu dents is not limited to lan-

guage skills. Figure 1.7 shows the distribution of student scores on the Mathematics portion of the CST. ALA students perform worse than EO students on the math CST, though by a smaller margin than on the ELA CST.

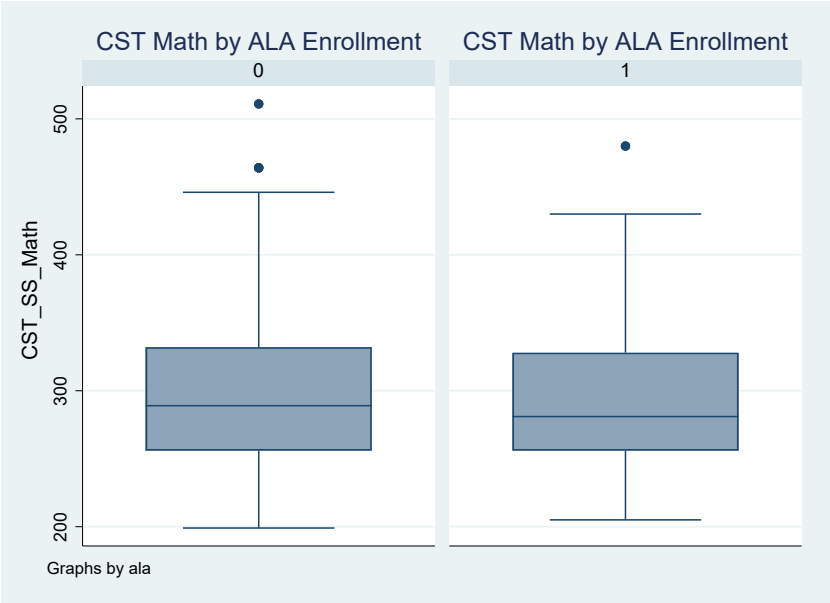


Figure 1.7: CST ELA Performance by ALA Enrollment

Enrollment in ALA has long-term effects on academic achievement. Figure 1.8 shows the distribution of Middle School grade point average (GPA) for ALA and EO students. Students enrolled in ALA during elementary school achieve significantly lower grades in Middle School than students enrolled in EO.

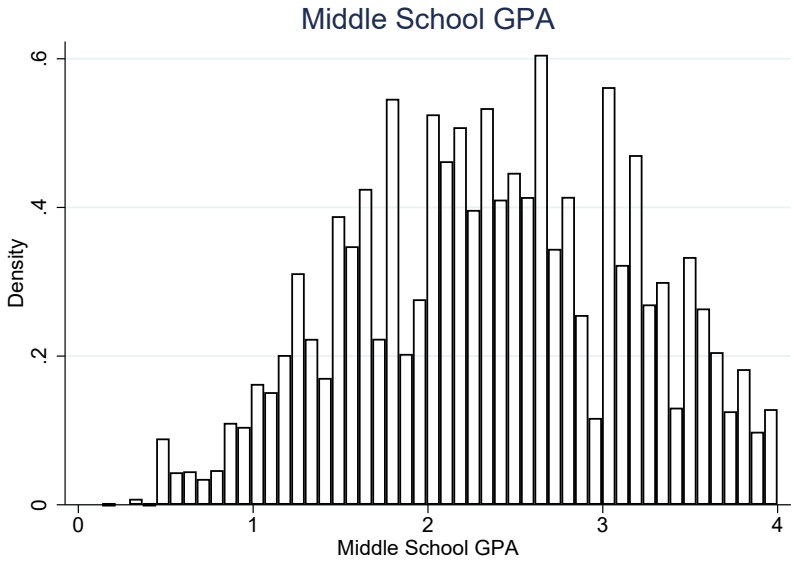


Figure 1.8: Distribution of Middle School GPA

Figures 1.9 and 1.10 show the correlation between CELDT score and ELA and math CST scores. CELDT score is positively correlated with both CST math and ELA scores, particularly at higher CELDT scores.

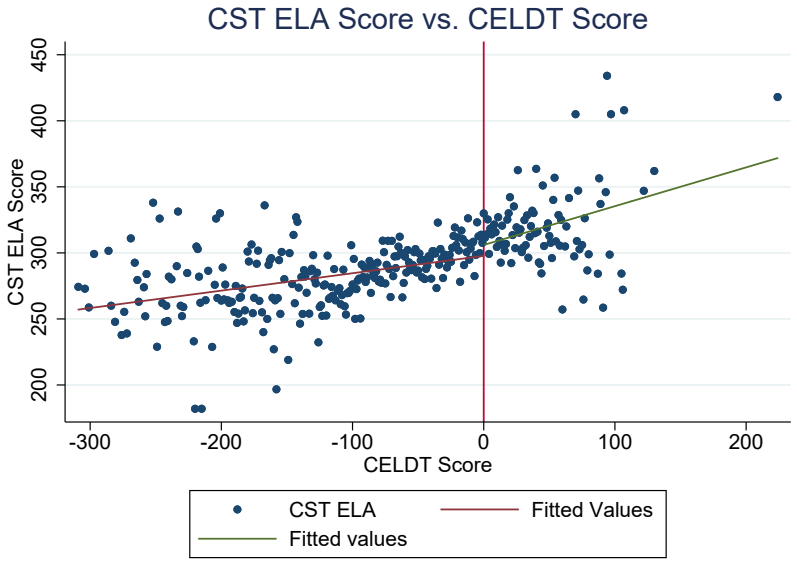


Figure 1.9: CST ELA Score by CELDT Score

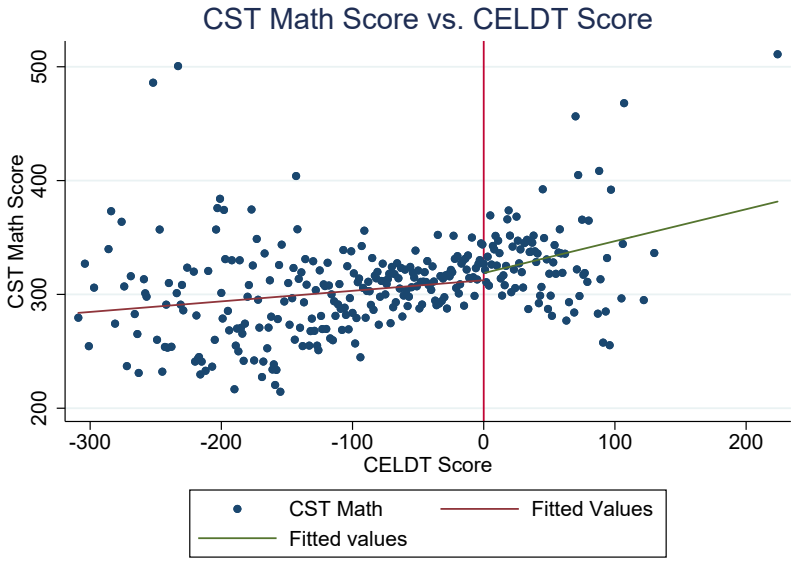


Figure 1.10: CST Math Score by CELDT Score

Figure 1.11 shows the correlation between CELDT score and middle school

GPA. There is a significant positive correlation between CELDT score middle school GPA, particularly at high CELDT performance levels.

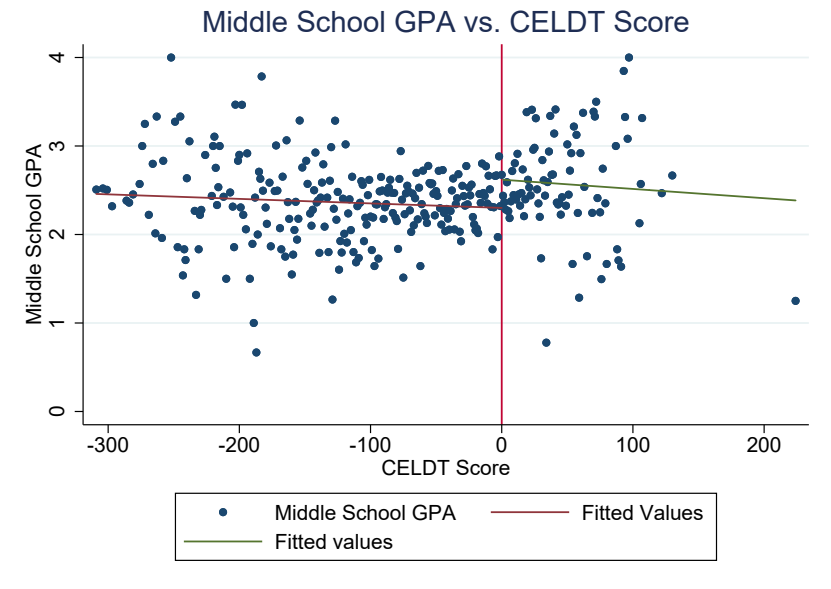


Figure 1.11: Middle School GPA by CELDT Score

Lower standardized test scores and middle school grades for ALA students has significant implications for long-term academic achievement and college-going. As shown in Figures 1.12 and 1.13, middle school GPA is highly correlated with both the verbal and math components of the Preliminary SAT, also known as the National Merit Scholarship Qualifying Test (NMSQT). Students take the PSAT in preparation for the SAT. A student's score on the PSAT can be an indicator of how they will perform on the SAT, and may therefore influence a student's decision to take the SAT and apply to colleges. ALA students have significantly lower middle school GPA, which decreases their probability of achieving a high score on the PSAT.

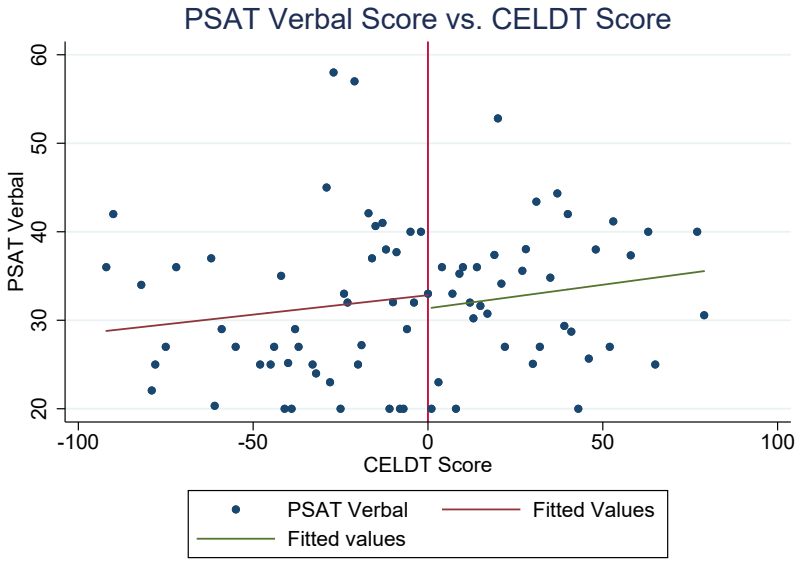


Figure 1.12: PSAT Verbal Score by CELDT Score

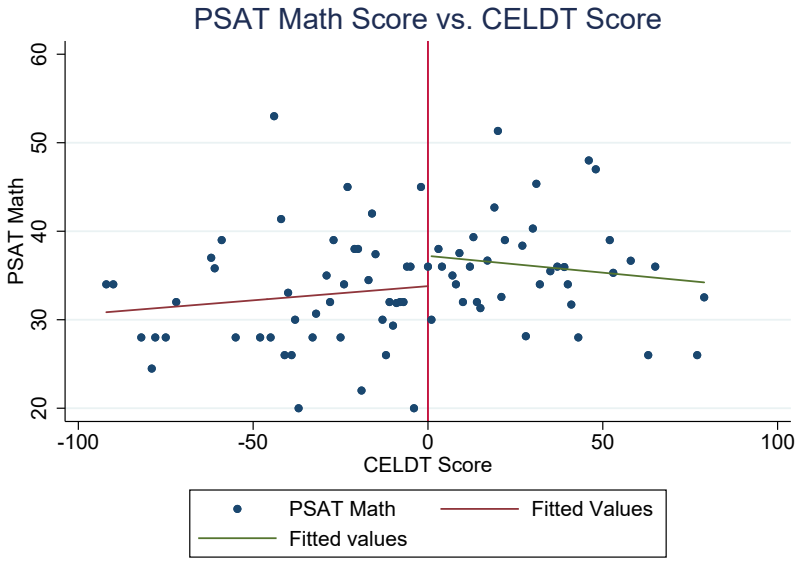


Figure 1.13: PSAT Math Score by CELDT Score

1.4.2 Data Limitations

There are several potential difficulties with estimating the causal effect of ELL programs using the NCSD data. First, the administrative data on enrollment in ALA and EO programs does not directly correspond to course enrollment data in ELL programs, and does not provide year of enrollment. I reconcile this by using the direct course enrollment data to determine whether an individual student was enrolled in an ELL program in a particular year and grade level.

Second, the initial administration of the CELDT is given based on responses on a parent survey to whether or not English is the students home language. Parents could have an incentive to state that English is spoken in the home even if it is not in order to ensure their childs initial classification into an EO classroom. Although teacher identification of a student as a non-native English speaker is also used in deciding whether or not to administer the CELDT to a student, it is possible that some non native English speakers are not taking the CELDT due to intervention by parents or misinformation on the home language survey. This would bias results as this parental involvement and motivation could be a confounding factor driving later academic success, and thus my results would over-estimate the negative effect of the program on student achievement.

Third, the CELDT is scored at both the local and state level. Because the local score is available immediately while it can take several months for the state score to become available to the district and school, initial placement is based on the local score. However, the official score of record is the state score, and we use the state scores

in this paper. This could create bias if there are differences between the state and local scores, particularly if the state scores are consistently higher or lower than the local scores. However, the state and local scores are almost always equivalent. Nonetheless, to overcome this bias, I use local scores.

Fourth, student take the CELDT when they first enroll in the district, and each year until they are reclassified as English proficient. This generates both initial and annual scores. Initial score determines initial eligibility for ALA, and probability of assignment to an ALA or EO classroom. Annual years in consecutive years must be above the English proficiency threshold in order for a student to be reclassified as English proficient, but a score above the CELDT threshold level does not guarantee reclassification as English proficient. Students must also score highly on the CEST ELA and pass a writing performance based assessment. At the same time, students enrolled in ALA for several years have different exposure to language development resources than students initially enrolled in an EO classroom. To overcome this endogeneity, I consider only local initial scores for kindergarten students taking the CELDT for the first time upon their initial enrollment in the district.

Fifth, some schools do not offer ALA. Most schools that offer ALA have more than half of their students enrolled in ALA programs. I use variation in ALA enrollment between schools in order to isolate the effect of ALA on academic outcomes. However, enrollment in an ALA or non-ALA school is not entirely exogenous. Parents of ELL students at schools without ALA programs can sign a parental exemption waiver in order to have their child transferred to a school that offers ALA. However, I am able

to observe assigned school and enrolled school, and verify whether there are systematic differences between transfer students and non-transfer students.

1.5 Model

I use a fuzzy regression discontinuity design (RDD) to estimate the causal effect of enrollment in ELL programs on student achievement. RDD analysis relies on discontinuity in the probability that a student is enrolled in a particular language development program based on whether he or she scores above or below a threshold score for English proficiency. The score threshold is used as an instrumental variable for treatment in order to test whether students enrolled in ALA have different academic outcomes than their observably identical peers enrolled in EO. The use of a regression discontinuity design assumes that students enrolled in ALA and SEI would experience the same trajectory in language abilities and academic achievement absence the treatment. Under this assumption, assignment to ALA or EO is essentially random between individuals who score just below the CELDT score threshold and those who score just above the threshold. Future work modify this model to include a Multiple Rating Score Regression Discontinuity (MRSRD). [65]

RDD studies are based upon the conditional expectation function (CEF) of the outcome variable. [11] In this study, this is various measures of academic achievement, such as middle school GPA, CST math or language score, or PSAT score, given the variable of interest, in this case CELDT score, represented as $E[Y_i|S_i = S]$. The

causal effect of interest is $f_i(S)$, the effect of score on academic outcome, conditional on enrollment in ALA or SEI. The difference in academic outcome between individuals with Score S* with score S*-1 is given by:

$$\begin{aligned} E[Y_i|S_i = S] - E[Yi|Si = S-1] = E[fi(S)fi(S-1)|Si = S] \\ +E[fi(S-1)|Si = S]E[fi(S-1)|Si = S-1] \end{aligned} \tag{1.1}$$

The first term is the decomposition of the average causal effect of going from a score of S-1 to S on the CELDT. However, the counterfactual average $E[f_i(S-1)|S_i = S]$ is not observed. The second term is the difference between the average academic outcome of those students with score S-1 and the academic outcome of those with score S, or the omitted variable bias in this model. Selection on observables states that:

$$E[f_i(S1)|X_i, S_i = A] = E[f_i(S1)|X_i, S_i = S1] \tag{1.2}$$

for all S, so that selection on observables eliminates omitted variable bias, and conditional on a set of observable characteristics Xi, the CEF and average causal response function are the same:

$$E[Y_i|X_i, S_i = S] = E[f_i(S)] \tag{1.3}$$

Then the causal effect of CELDT score is estimated conditional on X:

$$E[Y_i|X_i, S_i = S] - E[Y_i|X_i, S_i = S-1] = E[f_i(S)f_i(S-1)|X_i]. \tag{1.4}$$

In this paper, eligibility for enrollment in ALA changes abruptly at the threshold score for English proficiency on the CELDT, which is varies with grade, subject area and year. Because this rule is not followed precisely and depends upon the availability of ALA programs at each school, this can be interpreted as a fuzzy RDD in which the probability of treatment, i.e. enrollment in ALA, changes at the threshold CELDT score due to exogenous scoring rules that determine a students English proficiency level classification.

In conducting my econometric analysis, I follow the methodology of Card et al. (2008). [20] The relationship of interest is the effect of enrollment in ALA on the future academic outcome of individual i in score range s on CELDT subtest l, Y_{ilst}:

$$Y_{iklst} = \beta X_{is} + f(s_l) + \beta^P P_{ils} + u_{ilst} \tag{1.5}$$

where X_{is} is a vector of characteristics of individual in score range s, $f(s_l)$ is a polynomial representing the subset l score profile of the outcome, P_{ils} is an indicator equal to 1 if individual i in CELDT score range s on subtest l is enrolled ALA, and 0 otherwise, and u_{ilst} is an error term. Then β^P is the difference in outcome between an individual who is enrolled in ALA and one who is not.

The problem with estimating equation (1) is that program enrollment is en-

dogenous to language development, cognitive ability, parental involvement, and other observable and unobservable covariates. Because students are placed in ELL programs based on their English proficiency, the probability of enrollment in a given ELL program is discontinuous at the threshold CELDT score. As a result, the CELDT threshold for classification as LEP provides an instrumental variable for program enrollment.

This gives the first stage equation:

$$P_{ils} = \alpha X_{ils} + g(s_l) + \alpha^B B_{ilst} + v_{ilst} \tag{1.6}$$

where $g(s_l)$ is the subtest l CELDT score profile in enrollment in ALA, and B_{ilst} is an indicator equal to 1 if individual i in score range s scored below the threshold level for English proficiency on subtest l of the CELDT in year t.

Combining (1) and (2) gives the reduced form equation:

$$Y_{ilst} = \gamma X_{ils} + f(s_l) + \gamma^B B_{ilst} + e_{ilst} \tag{1.7}$$

where γ is a vector of coefficients, $f(s_l)$ is a score profile of outcome, and $\gamma_I V = \gamma^B/\alpha^C$ is the ratio of the reduced form coefficient of the threshold indicator B_{ilst} to the first stage coefficient.

Let P_{ils} and Y_{lst} represent population means of enrollment in ALA and academic outcome for individuals with CELDT subtest l score range s. Then (2) and (3) can be fitted to individual ALA enrollment, CELDT score, and academic outcome data using OLS and probit models, giving:

$$P_{ls} = psi_1 + g(s_l) + \beta^P B_{lst} \tag{1.8}$$

and

$$Y_{ls} = psi_2 + f(s_l) + \gamma^B B_{lst} \tag{1.9}$$

Because CELDT score is measured continuously, β^P and γ^P can be estimated by taking the average of P and y over score ranges on either side of the threshold.

1.6 Analysis

I limit my analysis to students who take the CELDT in elementary school and, in particular, in kindergarten. Although I have data on students in preschool through adult education, including elementary, middle and high school, I focus on elementary school students and on students who first enter NCSD and take the CELDT in kindergarten. This is because initial classification as limited English proficient (LEP) is dependent on initial CELDT score, while reclassification as English proficient depends on a later administration of the CELDT on an annual basis until a student achieves proficiency. CELDT scores can be either an annual score or the initial score upon the first administration of the CELDT to the student. Although I show results for all score types, I limit my analysis to initial scores for each studentin my preferred specification. Once a student reaches middle or high school, the course options available to them based

on their English proficiency change and vary by school and past exposure to language development programs.

The CELDT is scored both locally when a student takes the exam and by the state after test results are sent in for analysis. Because state scoring can take several months and the student must be placed in a class upon enrollment in the district, local scoring is generally used to initially place a student in a classroom. Although state and local scores are almost always equivalent, it is the local score that should be directly correlated with assignment to ALA or EO classroom For this reason, while I show results for all score types, I limit my analysis to local scores. Scores can also be from a school or district from which the student transferred. Because there is variation in these scores, I ignore transfer scores and use local scores in my preferred specification.

Although the probability of enrollment in ALA is significantly negatively correlated with CELDT score, the first stage is not as clean as I would like. Table 1.6 shows the results of a probit of ALA enrollment on CELDT total score and demographic controls for whether a student is Hispanic, Spanish-speaking, and of low socio-economic status (LSES). Results are given for all students, and for separately for students at schools with over 50% of students enrolled in ALA. Regardless of proportion of students at the school enrolled in ALA, CELDT score is significantly negatively correlated with probability of enrollment in ALA. Hispanic ethnicity, Spanish native language, and LSES are all significantly positively correlated with probability of enrollment in ALA.

Table 1.6: Probability of ALA Enrollment

	ALA Enrollment			
	All	All	High	High
Total Score	-0.0059^{***}	-0.0050^{***}	-0.0058^{***}	-0.0059^{***}
	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Hispanic	-	0.8617^{***}	-	0.9482^{***}
	-	(0.1057)	-	(0.1370)
Spanish	-	1.3927^{***}	-	1.4196^{***}
	-	(0.1063)	-	(0.1268)
LSES	-	0.7862^{***}	-	0.4633^{***}
	-	(0.0586)	-	(0.0870)
n	10,778	10,778	5,566	5,566

Table 1.7: First Stage

Figure 1.14 provides a graphical depiction of the first stage results for all students and grades in all years, using the local initial total CELDT score. There is a significant negative correlation between CELDT score and probability of ALA enrollment for most score levels. The lowest CELDT scores correspond to a probability of ALA enrollment of approximately 0.80. This suggests that nearly a quarter of the lowest performing students in terms of language ability are not enrolled in ALA. For students scoring above the threshold level for English proficiency, nearly all students are enrolled in EO.

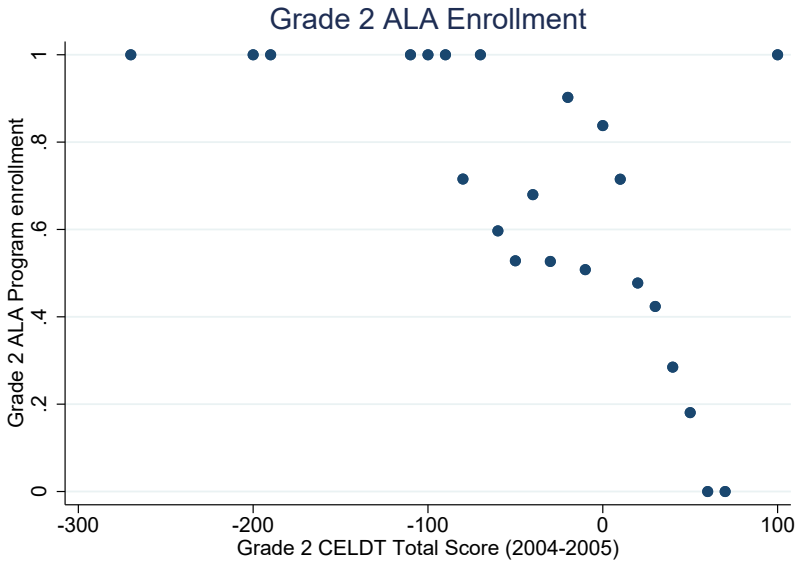


Figure 1.14: First Stage - All Students

Figure 1.15 provides that same first stage results for only kindergarten students. The results are nearly identical. This is most likely because the vast majority of students with an initial CELDT score, who are taking the CELDT for the first time, are kindergarten students. Nonetheless, the figure provides evidence that there is not significant variation across grade levels in probability of assignment to ALA based on CELDT score.

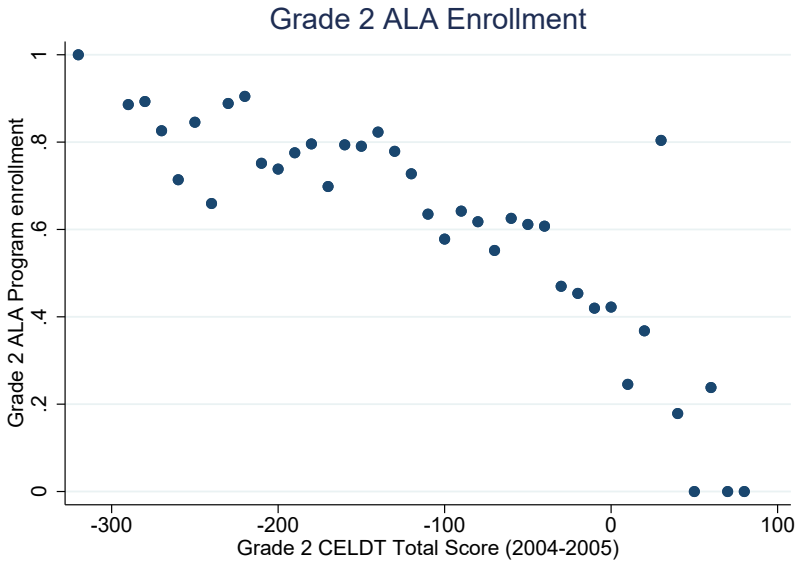


Figure 1.15: First Stage - Kindergarten Students

Figure 1.16 provides the same first stage results for kindergarten students in a single academic year, 2008-2009. Restricting to the 2008-2009 academic year, the negative correlation between CELDT score and ALA enrollment probability appears more linear. There is some evidence of discontinuity in probability of ALA enrollment at the threshold for English proficiency, though this discontinuity is small, less than 0.10. The 2008-2009 results further suggest that for some years and sub-groups, all students scoring at the lowest CELDT levels are enrolled in ALA.

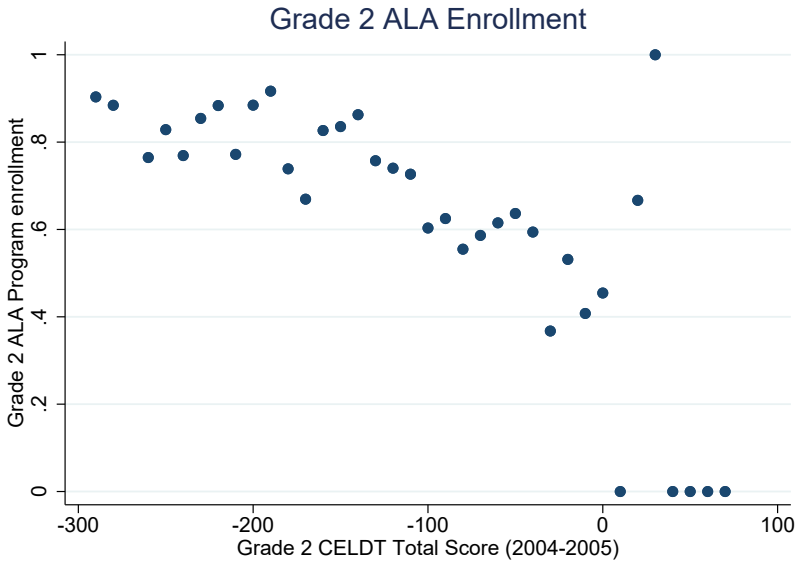


Figure 1.16: First Stage - Kindergarten Students, AY2008-2009

Unfortunately, for most years and subgroups, schools do not appear to be observing the rule for assignment to ALA classroom based on CELDT score. Some students scoring above the CELDT threshold are enrolled in ALA, while even for students scoring significantly below the threshold, less than half are enrolled in ALA.

This may be due to the fact that not all schools in the district have ALA classes available. Therefore I restrict my analysis to schools in which ALA classes are available. I choose several thresholds for proportion of students enrolled in ALA to define high vs. low ALA schools, and my results are robust to these thresholds. My preferred specification simply defines any school that offers ALA as a high ALA school (the lowest proportion of students at any school that offers ALA is 0.20).

Figure 1.17 shows ALA enrollment by CELDT score for kindergarten students

at schools offering ALA programs. Again, there is some evidence of discontinuity at the threshold for English proficiency. However, even at these school in which ALA is available, almost half of students scoring 100 points below the threshold are not enrolled in ALA. This could be due to limited availability of ALA classrooms of certified bilingual teachers, or due to parental decisions to place their children in EO classrooms. I am not able to disentangle these from the data that I have.

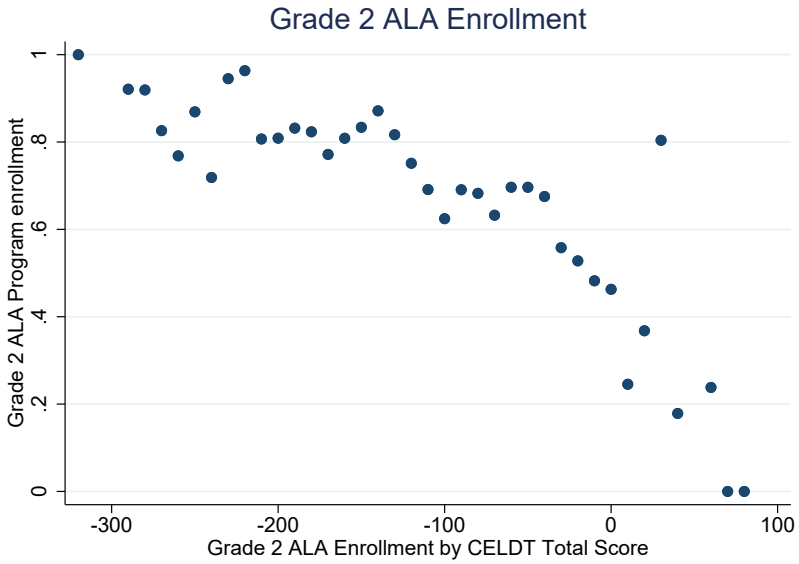


Figure 1.17: First Stage - ALA Schools

Figure 1.18 shows equivalent results for the academic year 2007-2008. Again, there is some evidence of discontinuity at the threshold score that could potentially be leveraged in an RDD.

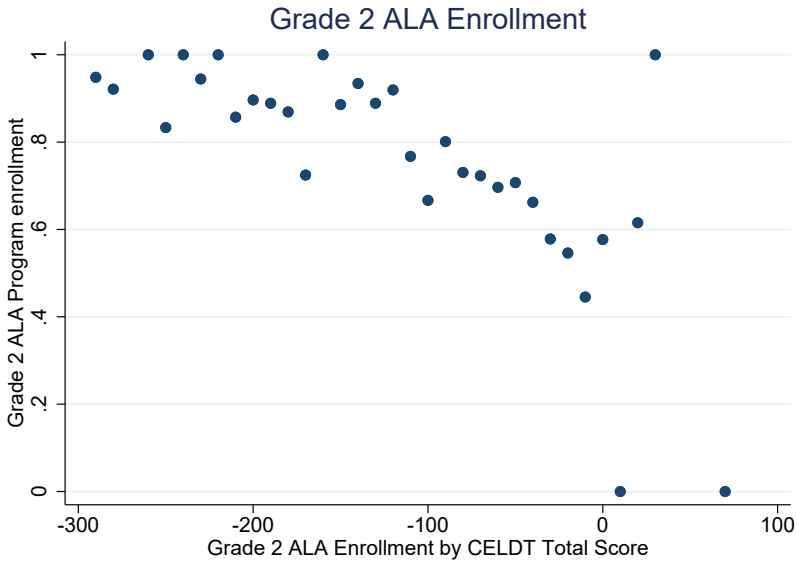


Figure 1.18: First Stage - ALA Schools, AY2007-2008

Figure 1.19 shows the first stage results using the listening subtest of the CELDT for kindergarten students at schools with ALA programs for all academic years, 2004-2010.

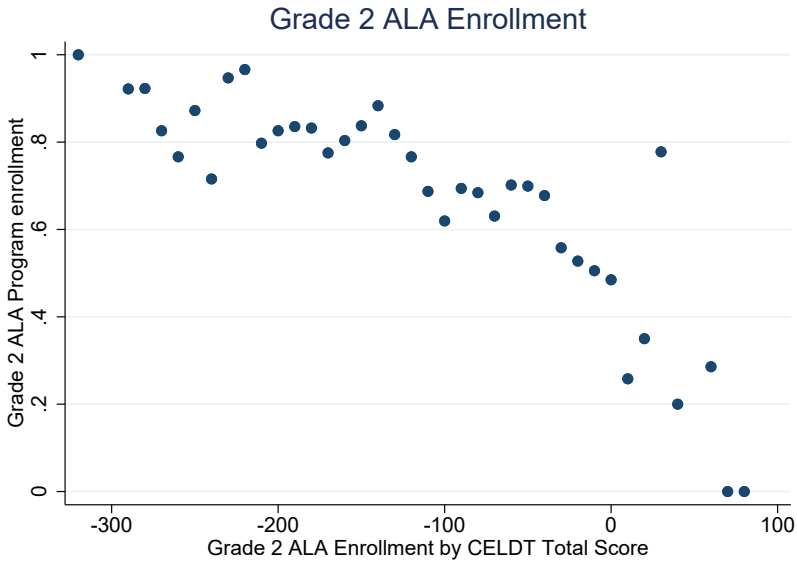


Figure 1.19: First Stage - CELDT Listening Subtest

Figure 1.20 shows the equivalent results for the speaking subtest. It does not appear that either of these subtests is the limiting factor in students' passing the CELDT. There is some evidence for higher grades that the writing subtest score is binding in determining whether or not a student passes the CELDT. However, Kindergarten and first grade students do not take the reading or writing subtests of the CELDT.

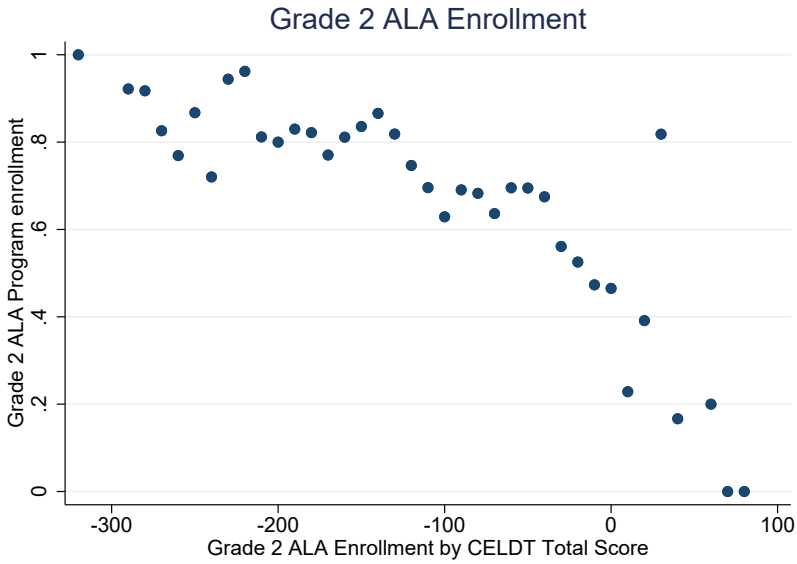


Figure 1.20: First Stage - CELDT Speaking Subtest

Unfortunately, due to the lack of compliance with the assignment to ALA rule based on CELDT score, my first stage results are not strong enough to achieve meaningful second stage estimates. Nonetheless, I provide a preliminary second stage analysis in the hopes that further data manipulation and understanding of program implementation will strengthen my first stage results.

The following results are for the effect of ALA on Middle School GPA and CST ELA and Math scores for students who took the CELDT in Kindergarten at schools offering ALA, based on local initial CELDT score. These figures are produced using the traditional RDD:

$$Y_i = \mu_1(Score_{ilst} - S^*_{lt}) + \mu_2(Score_{ilst} - S^*_{lt})^2 + \mu_3 Below_{ilst} + \\ \mu_4(Score_{ilst} - S^*_{lt})Below_{ilst} + \mu_5(Score_{ilst} - S^*_{lt})^2 Below_{ilst} \tag{1.10}$$

where Y_i is the outcome of interest for individual i, S^*_{lt} is the threshold score for English proficiency on section l of the CELDT in year t, and $Below_{ilst}$ is an indicator equal to 1 if the individual scored below S^*_{lt} on section l of the CELDT in year t. The coefficient of interest is μ_3, the second stage coefficient of the effect of scoring above the CELDT threshold for English proficiency on academic outcomes Y_i.

Table 1.8 gives the results based on CELDT total, listening, speaking, reading and writing local initial scores for students taking the CELDT in Kindergarten at schools offering ALA. There is a significant negative effect on CST ELA score of scoring below the CELDT threshold for English proficiency. However, there is no evidence of an effect on CST Math score or Middle School GPA.

Table 1.8: CELDT RD - CST ELA & Math Scores and Middle School GPA

	CST Verbal Score	CST Math Score	Middle School GPA
Total	−9.67**	0.07	0.05
	(106.71)	(114.99)	(0.04)
Listening	−137.64	−54.34	-0.04
	(89.77)	(111.95)	(0.20)
Speaking	63.95	−7.91	0.23
	(140.59)	(153.23)	(0.20)
Reading	132.02	0.21	-0.06
	(179.20)	(160.29)	(0.70)
Writing	−163.21	102.06	−0.29**
	(190.62)	(163.68)	(0.15)
Obs.	6,516	6,516	3,151

Figure 1.21 provides a graphical representation of these results for CST ELA score. There is some evidence of discontinuity in CST Verbal Score at the CELDT threshold score. This would suggest that students enrolled in ALA due worse in terms of English language skills and development than students of similar characteristics who are enrolled in EO.

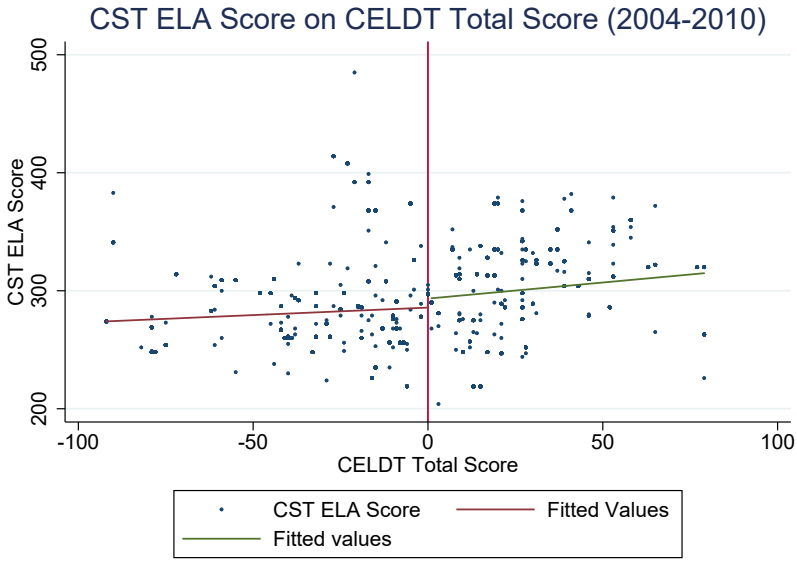


Figure 1.21: CELDT RD - CST Verbal Score

Figure 1.22 provides a graphical representation of these results for CST Math score. Although there is a positive correlation between CELDT score and CST math score, there is no evidence of discontinuity at the threshold score for English proficiency.

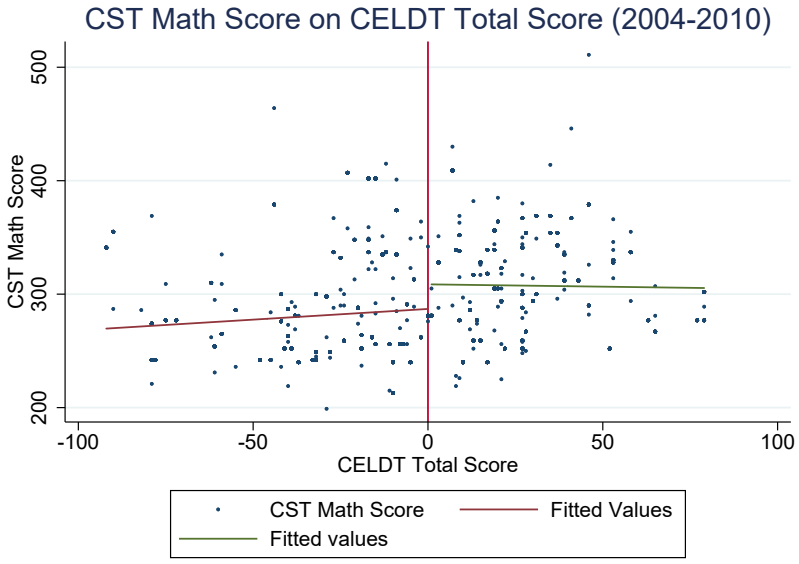


Figure 1.22: CELDT RD - CST Math Score

Figure 1.23 provides a graphical representation of these results for Middle School GPA. There is no evidence correlation between CELDT score and Middle School GPA, and no evidence of a sharp discontinuity at the CELDT proficiency threshold.

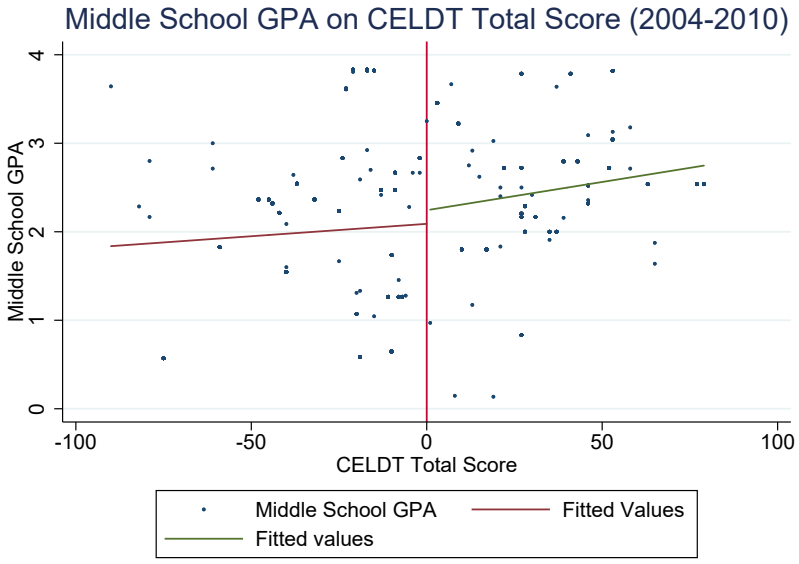


Figure 1.23: CELDT RD - Middle School GPA

In order to observe whether there is an explanation for the lack of first stage, in Table 1.9 I track several Hispanic, Spanish-speaking, low socioeconomic status students who scored below the English proficiency threshold on the CELDT in kindergarten at schools that offer ALA programs. Student 1 is observed in kindergarten/first grade in the 2004-2005 academic year, and in Grades 2 and 3 in the following academic year, at the same school. He persistently scores at Level 3, intermediate, on the CELDT. This may be an example of placing students who fail the CELDT only marginally in mainstream EO classrooms. Student 2 is observed in kindergarten in the 2004-2005 academic year, then in kindergarten/first grade and Grade 2 in the following academic years. He performs at the lowest level on the CELDT in kindergarten, then scores at level 4 and level 5 in the following years. This may be evidence that teachers or parents make

discretionary decisions regarding students' placement in non-ALA classrooms, particularly for younger, kindergarten students. Student 3 is observed in kindergarten through Grade 4 in consecutive academic years, with widely varying CELDT scores. Student 4 is observed in non-ALA kindergarten in the 2004-2005 academic year, then in bilingual ALA classrooms in Grades 1 and 2 in the following years, with low and intermediate CELDT scores. Student 5 is observed in kindergarten through Grade 4 in consecutive academic years at the same school, with steadily increasing CELDT scores. Student 6 appears to begin kindergarten in an EO classroom, and then transition to an ALA kindergarten classroom in the following year. Student 7 goes through EO kindergarten to Grade 2 in consecutive academic years at the same school, with steadily increasing CELDT scores. Finally, Student 8 is observed in kindergarten in the 2005-2006 academic year at an intermediate CELDT level, then in a Grade 5 bilingual classroom; this may provide evidence that students who just fail the CELDT are initially placed in EO classrooms, and then transitioned to ALA classrooms if they perform poorly in the mainstream EO classrooms.

Table 1.9: Student Course Enrollment

Student	Academic Year	School	Course	CELDT Level	ALA
1	2004-2005	339	GRADE K-1	3	0
1	2005-2006	339	GRADE 2	3	0
1	2006-2007	339	GRADE 3	3	0
2	2004-2005	317	KINDERGN AM	1	0
2	2005-2006	317	GRADE K-1	4	0
2	2007-2008	317	GRADE 2	5	0
3	2004-2005	317	KINDERGN AM	1	0
3	2005-2006	317	GRADE 1	5	0
3	2006-2007	317	GRADE 2	1	0
3	2007-2008	317	GRADE 3-4	3	0
4	2004-2005	313	KINDERGN AM	2	0
4	2005-2006	313	GRADE 1 BIL	3	1
4	2006-2007	313	GRADE 2 BIL	3	1
5	2004-2005	317	KINDERGN AM	2	0
5	2005-2006	317	GRADE K-1	4	0
5	2006-2007	317	GRADE 2	4	0
5	2007-2008	317	GRADE 3-4	5	0
6	2004-2005	354	KINDERGN AM	1	0
6	2005-2006	354	KINDR AM BIL	1	1
7	2005-2006	313	KINDERGN AM	2	0
7	2006-2007	313	GRADE 1	3	0
7	2007-2008	313	GRADE 2	5	0
8	2005-2006	355	KINDERGN AM	3	0
8	2009-2010	339	GRADE 5 BIL	3	1

Without a strong first stage, it is impossible to draw meaningful conclusions concerning the effect of bilingual education programs such as ALA on student achievement. Schools do not appear to be following the assignment to treatment rule. Analysis at the school level does not improve the first stage results.

1.7 Conclusion

My paper uses a new, longitudinal student dataset from a large school district in northern California. To my knowledge, my paper is the first to empirically consider the long-term academic outcomes of students initially enrolled in English Only versus bilingual programs at the district level in California. I find evidence that students enrolled in bilingual programs do worse than their peers enrolled in mainstream English only classrooms. My findings suggest that ELL programs may be detrimental to the long-run academic success of non-English speakers and may perpetuate rather than decrease the achievement gap between White and Hispanic students in California and the U.S. However, the statistical significance of my results are undermined by the systematic violation of the assignment to treatment rule across the district.

Because there are inertial effects determining a students path through the public school system, initial placement upon entry into the school district, normally in kindergarten, has significant and long-term effects on the classes a student is enrolled in and the peers he or she interacts with. Initial classification as limited English proficient (LEP) and enrollment in ALA has significant and long-term effects on academic outcomes.Students initially enrolled in ALA have a high probability of remaining in ELL programs throughout elementary school and into middle and high school, and achieve lower grades and standardized test scores than their peers in EO programs. These results have several significant policy implications.

Bilingual education programs are a highly contentious issue in California and

across the United States, with significant long-term effects for student achievement, the equitable distribution of education resources, and the development of human capital, all of which contribute to long-term inequality in career and wealth outcomes between native and non-native English speakers. Proponents of ALA argue that students enrolled in ALA will be able to maintain their culture and roots in their native language, while gradually being assimilated into the mainstream English speaking culture. However, many students enrolled in ALA remain in the ALA track throughout elementary school, and never transition into a mainstream English classroom. These students English listening, speaking, reading, and writing skills may never achieve a level of proficiency. As a result, upon reaching middle school, these students are put into ELD or SDAIE classrooms with their friends and peers, and can stay in these non-mainstream English classrooms throughout middle and high school. For example, Although California high school students are required to pass the California High School Exit Exam (CAHSEE) in order to receive a high school diploma, very few English learners pass the exam. In 2004, only 48% of the Class of 2004, and only 19% of ELLs, had passed the exam. [35] This has the potential to contribute to racial and linguistic segregation in school districts, as well as long-term educational and socioeconomic inequality.

However, in both Calfornia and the United States, public education policy must address the needs not only of ELLs, but also of English speakers. While integrating ELLs into mainstream English classrooms may have positive benefits for ELLs, it may be detrimental to English speakers. This raises significant and politically contentious issues of social equity both for elementary school students in classrooms today, and for

society as a whole as these students develop the learning resources and human capital that will propel them throughout their entire lives.

Public education policy pertaining to bilingual education and ELL students is becoming increasingly important as the number of ELL students in the public school system grows. With increasing number of immigrants in the U.S. school system, the question of how children acquire English is at the forefront of national education policy. Given its importance in classification of ELLs, the CELDT requires constant modification and updating. Assembly Bill 124, which was signed into law in October of 2011, requires that Californias State Superintendent of Public Instruction establish a committee of experts in English language to aid in the revision and updating of California ELD standards, from which the CELDT is developed. These standards were submitted and adopted in August and September of 2012, respectively. Alignment of the CELDT with the new 2012 standards is contingent upon state funding, which was absent in 2012-2013. However, the California Department of Education has taken action to proceed with CELDT updating once funds become available.

However, revision of the CELDT alone is not enough to ensure success for ELLs; the programs into which these students are placed as a result of CELDT classification must also be analyzed for the costs and benefits that they provide to students.Evidence of the long-term effects of ELL programs such as ALA for both ELLs and native English speakers has the potential to inform policy in developing educational programs that benefit students of all levels of English proficiency.

1.8 Future Research

Although the first stage in this RDD analysis of assignment to ALA based on CELDT score is lacking, preventing my ability to draw meangingful conclusions about the long-term effects of ALA on student achievement, I plan to pursue my research on bilingual and other types of education programs. Future research will continue in the area of education and English language development.

There are several potential extensions using the NCSD data set. The ALA program only involves elementary school students. Middle and High School students are who are classified as LEP are enrolled in ELD and SDAIE, respectively. Although analysis of these programs is subject to endogeneity bias due to previous years in language development programs, an analysis of the efficacy of these programs provides one avenue for extension using this school data. Additionally, several papers have looked into the effect of reclassification as English proficient on student outcomes. The timing of reclassification may have a significant effect on student achievement. [63] However, the sample size is significantly smaller and suffers from issues of endogeneity given that students have already been tracked into separate classrooms and program upon their arrival in the district and initial taking of the CELDT.

Chapter 2

Consumer Preferences for Safety in the New Vehicle Purchasing Decision

2.1 Introduction

Each year in the U.S., 37,000 people die in automobile accidents, with an additional 2.35 million injured or disabled. New vehicle fleet safety has increased over time due in part to safety technology improvements and policy intervention; in 1978, there were 50,133 motor vehicle fatalities in the U.S., despite lower vehicle miles traveled. Vehicle safety technologies have played a large role in this reduction. [47]

Current regulations governing vehicle safety include seatbelt laws, speed limits, and safety standards for crash prevention technologies, in addition to crashworthiness ratings such as the National Highway Transportation and Safety Administration (NHTSA)'s New Car Assessment Program (NCAP). Improvements in vehicle safety

technology are reflected in improvements in NCAP crash test performance over time. According to the Department of Transportation's Federal Register (2008), in 1979, the first year that vehicles were rated for frontal impact safety under NCAP, under 30% of vehicles received 4 or more stars (5 = most safe), whereas 98% of vehicle in model year (MY) 2007 were rated at 4 or 5 stars. [6] Does regulation of motor vehicles in the form of standardized safety ratings have a positive effect on safety?

Effective regulatory intervention of vehicle safety depends on how consumers make individual purchasing decisions. Empirical evidence suggests that vehicle safety is a primary concern for consumers. However, consumers face uncertainty in the vehicle purchasing decision, with imperfect information about vehicle safety. Further, consumers vary in their risk profiles and preferences for vehicle safety based on age, gender, income level, education, children, etc. [41]

Vehicle safety rating programs provide comparative information on vehicle safety to consumers. The two main vehicle safety rating programs in the US are NCAP and the Insurance Institute for Highway Safety (IIHS) ratings. The primary difference between NCAP and IIHS ratings is that NCAP evaluates crashworthiness, while IIHS evaluates both crashworthiness and crash avoidance. Both ratings are published in Consumer Reports and advertisements.

As of September 2007, all new vehicles sold in the U.S. are required to display NCAP safety ratings on the Monroney Sticker taped to the driver's side window. NCAP is a demand-side approach to improving safety. Using the discontinuity of the 5-star NCAP ratings in underlying crash test performance score and the September 2007

Monroney Sticker requirement as an event study, I evaluate consumer preference for safety in purchasing new vehicles and the efficacy of the NCAP program in promoting innovation in vehicle safety technology. I compare vehicles that just barely miss the nearest star threshold (e.g. 4-star vehicles) to stars that just barely exceed the nearest star threshold (e.g. 5-star vehicles). In theory, these vehicles should be identical in underlying attributes and probability of injury. However, due to the NCAP rating mechanism, some vehicles 'exogenously' receive an additional star in their safety rating. I evaluate the effect of receiving this additional star (relative to just missing the nearest star threshold) on vehicle sales.

If NCAP rating is a signal of true vehicle safety, and if consumers value vehicle safety, then we can expect to observe preference for vehicles that receive a higher NCAP star rating. However, previous literature suggests that producers may simply increase vehicle curb weight as a way to improve crash test performance and star rating. Further, producers may respond to an unfavorable NCAP rating by lowering price, investing in crash-prevention technology, or other compensating behavior. I do not find a sales response to receiving a higher NCAP rating. However, I do find evidence that vehicle manufacturers are responding to the NCAP ratings program by (1) increasing curb weight as a way to improve crash test performance and (2) raising price for vehicles that score above the NCAP star threshold.

The remainder of this paper is organized as follows: Section 2 reviews the literature and discusses my contribution. Section 3 discusses the U.S. National Highway Transportation and Safety Administration (NHTSA)'s New Car Assessment Program

(NCAP). Section 4 introduces the data and outlines my empirical strategy. Section 5 presents my results. Section 6 concludes.

2.2 Previous Literature

This paper evaluates consumers' willingness to pay product safety technology in the automobile industry, a market with significant public health and economic consequences. There is a large literature related to estimating willingness to pay by individuals for risk reduction. [45], [79] In the context of worker safety, for example, or consumer evaluation of risk of injury or death in the use of a particular product.

Dreyfus and Viscusi (1995) consider consumer price trade-offs for vehicle safety and fuel efficiency. They evaluate the effect of vehicle durability on consumer choice in order to evaluate an implicit rate of interest used by consumers in the new vehicle purchasing decision. They find that buyers have a distribution of preferences over vehicle attributes, but discount vehicle price for safety and fuel efficiency. [30]

Atkinson and Halverson (1990) estimate consumer willingness to pay for reductions in probability of death. [12] They use a hedonic pricing model in which vehicle performance characteristics such as safety technology directly enter the consumer utility function, while price and fuel efficiency affect the consumer's budget constraint. Hedonic models estimate vehicle price as a function of various vehicle attributes. The implicit marginal price of each attribute is the partial derivative of price with respect to that attribute. The authors model cost as a function of vehicle safety (predicted risk of death based on vehicle characteristics) and performance characteristics for each vehicle model. Real-world fatality rate is thus a function of a given model's safety and driver characteristics. Estimating this model for 112 vehicle models of MY 1978 vehicles, simi-

larly to Dreyfus and Viscusi, the authors of also find that buyers are willing to discount vehicle price for increased safety.

Berry, Levinsohn & Pakes (1995) evaluate whether the increased cost of safety features leads to a compensating price increase [15]. Their design endogenizes consumer choice over vehicle characteristics in modeling price and demand elasticity. Using an oligopolistic differentiated products model, the authors estimate cost as a function of vehicle characteristics and demand as an aggregation of individual consumer behavior in a discrete choice model. They find high demand elasticity for safety, and that the cost of safety features is passed through to the consumer via price.

Knittel et al. (2011) consider vehicle weight, horsepower, torque and fuel economy (CAFE) standards. They find only small gains in fleet fuel efficiency, largely due to increases in vehicle curb weight. He argues that vehicle manufacturers respond to the "arms race" over vehicle safety by increasing curb weight, at the expense of fuel efficiency. This race to the heaviest generates negative externalities on the road, particularly for occupants of lighter vehicle models. [48]

Do consumers respond to the mandatory disclosure of safety information via increased demand for safer vehicles? Consumers value product quality but have imperfect ex ante information about vehicle quality prior to purchase. Aggregated and standardized ratings decrease the cost of information acquisition and transfer. Learning about product quality involves a costly investment of time and resources for buyers, particularly for multi-attribute, technologically complex products such as automobiles. Ratings and reviews decrease information acquisition costs for buyers. Increased infor-

mation transparency over vehicle safety should, in theory, increase demand elasticity, resulting in increased intensity of price competition and decreased profits for producers. [8]

Safety is a key factor in the new vehicle purchasing decision for consumers, and previous studies have found a strong correlation between safety ratings and vehicle sales. Positive safety ratings significantly increase likelihood of purchase. [81] Safety is a primary factor in consumer choice over vehicle purchase. [49], [41], [80] Koppel et al. (2007) survey participants in Spain and Sweden about the new vehicle purchase decision. Respondents ranked safety over price, reliability, and other factors. The authors find that this preference is stronger among EuroNCAP (the European version of NCAP) users. [49]

Hellinga et al. (2007) evaluate the role of safety features in parents choice of vehicles for teenagers. In a 2006 survey of 300 parents in Minnesota, North Carolina, and Rhode Island, they find that parents rank safety, available family vehicle, and reliability as the top reasons for choosing a vehicle for their teenager to drive. [41]

Using a sample of 2,002 Canadian drivers' survey responses ranking importance of features that buyers consider when buying a car and running ANOVA tests, Vrkljan & Anaby (2011) find that safety and reliability are the highest ranked vehicle attributes among respondents, while design and performance are the lowest lowest. They find that younger drivers value safety less, with young male drivers rating safety lowest. [80]

Girasek & Taylor (2010) address the distributional and welfare effects of vehicle safety using correlations and ANOVA to estimate the relationship between socioeco-

nomic status and vehicle safety features. They link vehicle identification numbers and information on vehicle safety features with data on buyers' income and education. [38]

Empirical evidence shows a strong link between automobile safety and manufacturer reputation. The car market exhibits strong branding and advertising seen at the model level. Using data on 23 million vehicles registered in the UK between 1992 and 2002, Bates et al. (2007) find that automobile safety recalls can damage brand value and decrease stock price. [14] Choi & Lin (2008) evaluate consumer response to Mattel product recalls. [25] Rupp et al. 2002 evaluate damage to shareholders from firm vs. government initiated automotive recalls. [70]

However, ratings programs such as NCAP may have distortionary effects. Producers face a tradeoff between product attributes, such as safety, fuel efficiency, and performance. Vehicle manufacturers may improve crash test performance by increasing vehicle weight, at the expense of fuel efficiency and performance. Ito and Sallee (2014) evaluate corrective policy and "attribute-based" regulation by constructing a panel using a product identifier that is narrower than vehicle model. They find that producers respond to curb weight-based fuel efficiency standards by adjusting vehicle weight, exhibited in the data by bunching of vehicles at weight notches. This distortionary response by producers generates large welfare losses due to weight-related externalities. [43]

If manufacturers design new vehicles to perform well on NCAP crash tests in a simulated laboratory setting that differs significantly from real-world driving conditions, this will crowd out innovation in safety technology that saves lives in real-world crashes.

The automobile industry, particularly at the high end, exhibits technological innovation via product differentiation, rather than decreasing costs. Product variety is correlated with product durability/lifespan, and is higher than optimal unless an industry exhibits high fixed product launch costs and low substitutability (which is not the case in the car industry). [76] With high substitutability between differentiated products, the automobile industry exhibits high product market competition. Existing models of endogenous growth predict lower levels of innovation in more competitive markets. [29], [69], [39], [7]

2.3 The New Car Assessment Program

NHTSA's New Car Assessment Program provides comparative information on the safety of new vehicles to assist consumers with vehicle purchasing decisions and encourage motor vehicle manufacturers to make vehicle safety improvements. [Docket No. NHTSA-2015-0119] Its mission is to: (1) Help consumers with vehicle purchasing decisions, (2) Incentivize manufacturers to improve current safety performance and features of new vehicles; and (3) Promote innovation and development of new vehicle safety features. NCAP focuses on crashworthiness technology such as seatbelts and airbags, but also identifies whether rated vehicles are equipped with Crash Avoidance Technologies, such as Electronic Stability Control (ESC), Lane Departure Warning (LDW), and Forward Collision Warning (FCW). Additional crash avoidance technologies include rearview video systems (RVS), automatic emergency breaking, and tire pressure monitoring systems (TPMS).

NCAP was established 1978 under Title II of the Motor Vehicle Information and Cost Savings Act of 1972. In 1978, there were 50,133 motor vehicle fatalities in the U.S. NCAP and other transportation policies have successfully decreased motor vehicle deaths in the U.S.; in 2013, , there wer 32,719 motor vehicle fatalities in the U.S., despite increased vehicle miles traveled since 1978. NHTSA begain testing vehicles for crashworthiness using frontal driver and frontal passenger crash tests with 1979 model year. 5-Star ratings were introduced in 1994 for MY1990- vehicles. The side driver and side passenger crash tests were added for MY1997, and rollover ratings for MY2001.

Beginning in September of 2007, all new vehicles sold in the U.S. were required to display the Monroney Sticker, which provides the NCAP 5-star safety ratings, as well as fuel efficiency. Figure 2.1 shows the Monroney sticker for a 2010 Subarue Impreza. The 5-star NCAP ratings are provided for Frontal and side driver and passenger and rollover. Beginning with MY 2011-, the NCAP ratings formula was revised, with Overall Vehicle Score introduced as a weighted average of the frontal, side, and rollover test scores. Side pole testing was also introduced for MY 2011-. The Monroney sticker also provides information on fuel efficiency.

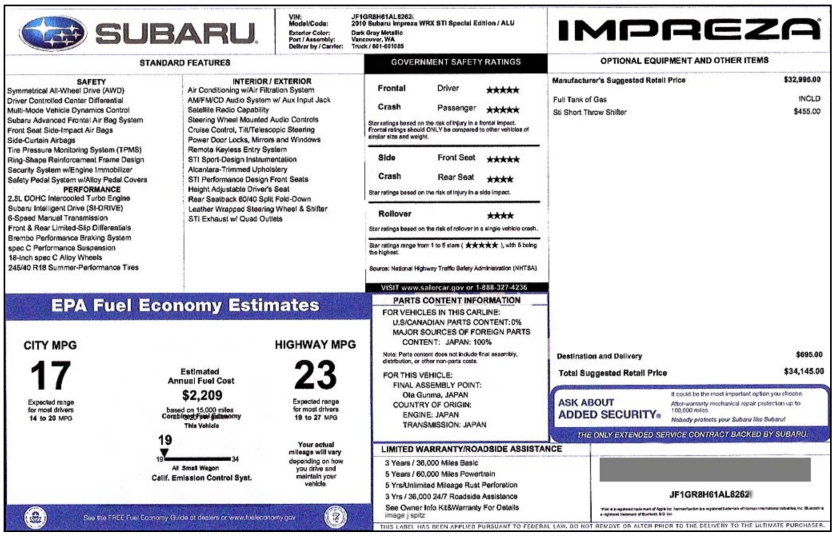


Figure 2.1: Monroney Sticker for 2010 Subaru Impreza

In selecting which vehicles to rate for each model year, NHTSA rates a sample of new vehicles that are predicted to have high sales volume, have been structurally redesigned, or have improved safety equipment. NHTSA selects vehicles for testing on the following basis: (1) provide NCAP ratings for all three programs (front, side, and rollover); (2) maximize the percentage of fleet coverage by selecting vehicle models with

high projected sales volumes; (3) focus on consumer or agency interests (e.g maximize percentage of fleet coverage); (4) examine new safety features; and (5) consider vehicle price. [6] Approximately 60% of the new light vehicle fleet for MY 2011 were tested by NHTSA, for example. NHTSA purchases tested vehicles from dealerships across the country; the vehicles are not supplied directly to NHTSA by the manufacturer. The NCAP rating is binding; it is carried over until a model is re-tested. Re-testing is done if an existing model is significantly structurally re-designed or if the manufacturer introduces significant safety technology improvements in an existing vehicle model.

NHTSA purchases vehicles from dealerships and evaluates their crash performance in controlled collisions, with risk of injury determined from sensors in crash test dummies. For example, in the Frontal Crash Test: Crash test dummies representing an average-sized adult male and a small-sized adult female are placed in the driver and front passenger seats, respectively, and are secured with seat belts. Vehicles are crashed into a fixed barrier at 35 miles per hour (mph), which is equivalent to a head-on collision between two similar vehicles each moving at 35 mph. Instruments measure the force of impact to each dummys head, neck, chest, pelvis, femur (legs), and feet. The frontal crash rating is an evaluation of injury to the head, neck, chest, and femur (legs) for the driver and right front seat passenger. Probability of injury is calculated via a continuous underlying vehicle safety score (VSS). Thresholds are then superimposed on the distribution of probability of injury to map VSS to a 5-star safety.

NCAP ratings are widely advertised by manufacturers. Cross-vehicle comparisons are regularly provided in mainstream press and consumer media (e.g. Consumer

Reports). Consumer reports and other consumer-targeted product reviews also include other ratings (e.g. IIHS, Edmunds), information on crash avoidance performance, fuel efficiency, etc. Manufacturers regularly advertise vehicles' test performance, and call attention to poor performance by a rival.

Although manufacturers run their own crash tests in-house, at the margin NCAP crash test performance is unknown by manufacturer. Further, test result is binding until re-testing is triggered by a significant change in model production. Producer choice over safety could be modeled as a repeated game, but I will argue that vehicle manufacturers face uncertainty over vehicle crash test performance and star rating.

NHTSA disseminates NCAP safety ratings via press release to over 1,000 organizations, including news services, consumer groups, magazines, etc., with readership in the tens of millions, including Consumer Reports, The Car Book (published by the Consumer Federation of America), and The Car Guide (published by the United States Automotive Association (USAA). Ratings are also published online. [42]

NCAP has become a de facto safety mandate in that manufacturers re-design poorly rated vehicles out of fear that consumers will deem them unsafe. In addition, manufacturers increasingly reference vehicle model and fleet safety features, including NCAP rating, in advertisements. [42]

2.4 Data & Methodology

Using data from the Department of Transportation Federal Registrar's Dockets, I construct a novel dataset matching continuous crash test performance measures to NCAP 5-star rating and monthly national sales at the model-level. I get information on model-level vehicle characteristics and safety technology from DataOne, which provides vehicle characteristics at the VIN-prefix level for the universe of light duty vehicles produced 1981-2016MY. I use monthly, model-level sales data from WARDS Automotive, sales and production data for the U.S. and North America at the subseries level 1980-2017. I use vehicle safety ratings from NHTSA's New Car Assessment Program. This includes crash test results and 5-star safety ratings for all rated vehicles from 1990-2016.-Crash test results and 5-Star Safety Ratings for all rated vehicles 1990-2016.

To generate a crosswalk between NCAP rating and DataOne, I match on lowercase make, model, body style, vehicle type, and drive train by calculating smallest Levenshtein distance. I then calculate the best NCAP match for each DataOne vehicle. For the majority of vehicles, once imposing a match on these characteristics there is a single possible match. For vehicles with multiple possible matches I calculate the standard deviation of NCAP ratings for each set of possible matches. In the majority of cases this is 0. Finally, I visually inspect ambiguous cases to hard-code a match in cases such as the BMW 128i/135i or the Lexus GS 350. In this way I am able to match 100% of the NCAP ratings to DataOne observations for MY1990-2010 vehicles.

Below are summary statistics and a balance test of covariates for all vehicles,

all test types, all vehicle types, all years, based on whether a vehicle scored above (+1 Star) or below (-1 Star) the rating threshold on a given crash test. Each observation represents one crash test. Vehicles that just make the nearest star rating threshold are significantly more likely to have a higher MSRP and curb weight, and lower fuel efficiency.

	+1 Star	-1 Star	Diff	
	mean	mean	b	p
MSRP	27461.46	24215.23	-3280.17***	0.00
Curb Weight	4223.99	3432.85	-799.75***	0.00
Height	67.34	60.63	-6.75***	0.00
MPG	18.90	20.21	1.28***	0.00
Max HP	193.51	166.60	-27.47***	0.00
Sales (volume)	1254.66	1849.26	597.95***	0.00
Sales (percent)	0.01	0.01	0.00***	0.00
Observations	195174	98582	293718	

NCAP distinguishes four vehicle types: Cars, Trucks, SUVs and Vans. Passenger Cars are further subdivided into mini, light, compact, medium and heavy based on curb weight. There is further information on vehicle body style, drive train, and production release. Ratings are provided at the vehicle make and model level. There are 5 types of crash test ratings: Frontal Driver, Frontal Passenger, Side Driver, Side Passenger, Rollover and Rollover 4WD. Table 2 provides summary statistics by vehicle type for Cars and Trucks, which represent the majority of vehicle models and sales.

	Cars	Trucks	Diff	
	mean	mean	b	p
MSRP	27461.46	24215.23	-2899.21***	0.00
Curb Weight	4223.99	3432.85	-1462.32***	0.00
Height	67.34	60.63	-15.04***	0.00
MPG	18.90	20.21	6.84***	0.00
Max HP	193.51	166.60	-7.66***	0.00
Sales (volume)	1254.66	1849.26	1626.52***	0.00
Sales (percent)	0.01	0.01	0.00	0.11
Observations	195174	98582	231233	

Using formulae from the DOT Federal Register [6], I am able to calculate the continuous risk of injury used by the NCAP program to generate discrete 5-star safety ratings from each crash test. I exploit discontinuity in the assignment of 5-star NCAP ratings across clearly defined star thresholds in order to estimate demand- and supply-side responses to just missing a star threshold. NCAP measures a continuous vehicle crash probability of injury risk. Nature of the testing process prevents manipulation along the running variable.

The Monroney Sticker reports discrete 5-Star Safety ratings for Frontal and Side Driver and Passenger and Rollover based on underlying probability of injury calculated from vehicle crash test performance. See Chapter 3 for the mapping formulae.

$$Stars = \begin{cases} 5, & \text{if } Pr(Injury) \leq T_{4,5} \\ 4, & \text{if } T_{4,5} \leq Pr(Injury) < T_{3,4} \\ 3, & \text{if } T_{3,4} \leq Pr(Injury) < T_{2,3} \\ 2, & \text{if } T_{2,3} \leq Pr(Injury) < T_{1,2} \\ 1, & \text{if } Pr(Injury) \geq T_{1,2} \end{cases}$$

Regression Discontinuity Design (RDD) analysis relies on several critical assumptions. First, assignment to treatment D_i is determined by the value of a particular covariate X_i being on either side of a threshold value c. The unconfoundedness assumption is thus given by: $Y_i(0), Yi(1)D_i|X_i$. This can be problematic because there are no values of X_i that overlap, requiring extrapolation in estimating the average treatment effect. Second, the monotonicity assumption requires that $D_i(X_i)$ be non-increasing in X_i at $X_i = c$. Finally, RDD analysis assumes continuity of the covariate distribution functions at the point of discontinuity (threshold).

Given the extensive amount of in-house testing conducted by vehicle manufacturers, one might expect car makers to make improvements to vehicles that seem likely to score below the 5-star probability-of-injury threshold. However, probability of injury is continuous across the ratings threshold, whereas gaming would be clear via bunching of test scores above the star threshold.

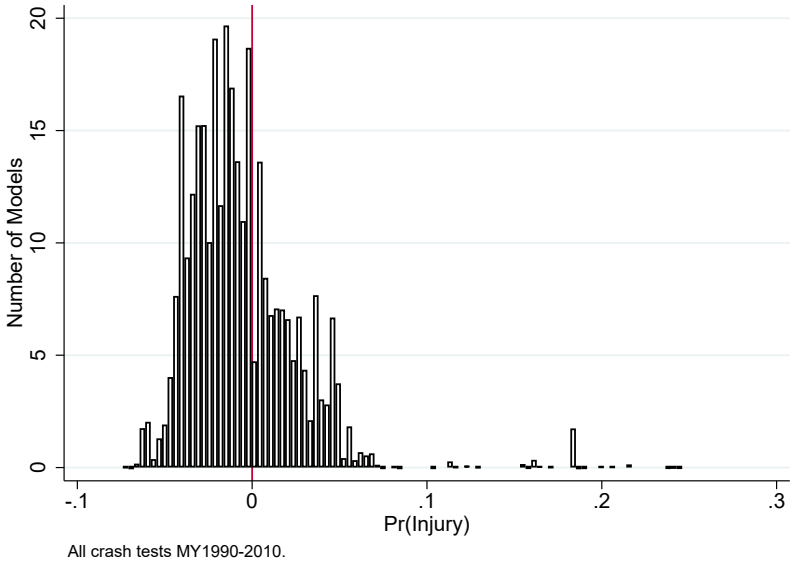


Figure 2.2: Density Test of Running Variable

NCAP provides a sharp RD design. Assignment to treatment $D_i = 1[X_i \geq c]$. Then $Pr(D_i = 1) = 1$ if $X_i \geq c$ and $Pr(D_i = 1) = 0$ if $X_i < c$.

The outcome of interest Y_i is evaluated via a binary treatment variable D_i indicating whether or not a vehicle scored at or above the nearest rating threshold T. Thus $D_i = 1P_{ist}(Injury) > T$ and let $Y_i(1)$ denote the outcome with treatment (i.e. vehicles that scored above the nearest star threshold) and $Y_i(0)$ denote the outcome without treatment (i.e. vehices that scored below the nearest star threshold.) Because each vehicle model receives only one star safety rating in each model year, the true effect of safety information disclosure $Y_i(1) - Y_i(0)$ is unobservable at the individual vehicle model level. Then the average treatment effect is $T = E[Y_i(1) - Y_i(0)|X_i = c] = E(Y_i(1)|X_i = c] - E(Y_i(0)|X_i = c]$.

Figure 2.3 shows the first stage, pooled for all test types and years. My running variable, continuous probability of injury, perfectly predicts star rating assignment, allowing for a sharp RD design.

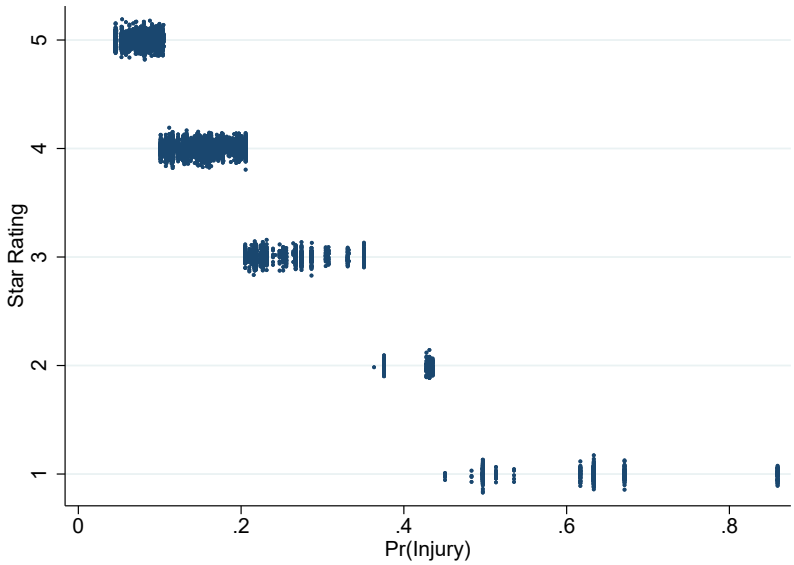


Figure 2.3: First Stage

To determine consumer demand for safety, I evaluate the correlation between Safety Rating and Sales:

$$Sales = \beta_0 + \beta_1 f(P(Inj)) + \beta_2(+1Star) + \beta_3 f(P(Inj))(+1Star) + \varepsilon \tag{2.1}$$

Pr(Injury) is calculated for each vehicle for frontal driver, frontal passenger, side driver, side passenger, and rollover. I use ratings and crash test data for MY1990-2010.

A fundamental limitation of RDD is that can only be used to the estimate effect

of treatment at point of discontinuity. This suggests a tradeoff between restricting the sample to be near the threshold, and including more data further from the threshold point. Further, the consistency and unbiasedness of RDD estimates depend on the absence of strategic behavior (gaming) at the threshold point.

2.5 Analysis & Results

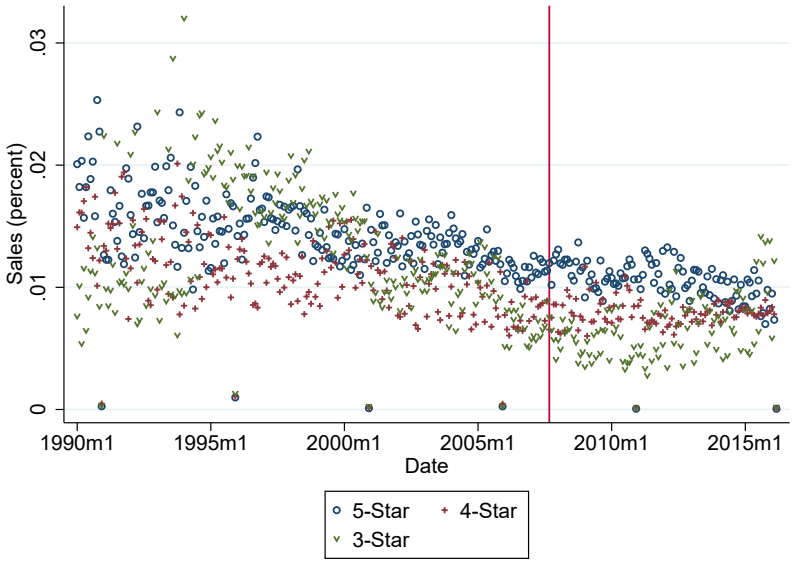
I measure consumer response to vehicle safety information disclosure as the percentage increase in vehicle sales at the model level. If consumers are responding to safety, we would expect to observe an increase in sales of vehicles scoring just above a star threshold relative to vehicles scoring just below a star threshold. I am interested in whether sales become more sensitive to safety as a result of the NCAP rating system Monroney Sticker.

The majority of models score between 3 and 5 stars. I therefore focus my analysis on the sales effect of scoring 5 (relative to 4) and 4 (relative to 3), i.e. estimating the treatment effect at the 5-4 and 4-3 star ratings thresholds. I report pooled results.

Because NCAP rates vehicles based on vehicle type, which is categorized based on vehicle curb weight, I investigate heterogeneity in sales response across vehicle class. I expect less of an effect of the NCAP in less competitive vehicle categories where consumers have limited choice over vehicle type.

If consumers value safety, then the 2007 Monroney Sticker requirement should, all else being equal, lead to an increase in sales of higher-rated vehicles. Figure 2.4 plots total series-level market share (in terms of monthly U.S. new vehicle sales) by star rating (1990-2015).

Figure 2.4: Sales



Figures 2.5, 3.5.1, and 2.7 show average series-level MSRP, curb weight, and fuel efficiency by star rating (1990-2015). The red vertical line delineates the September 2007 Monroney Sticker mandate.

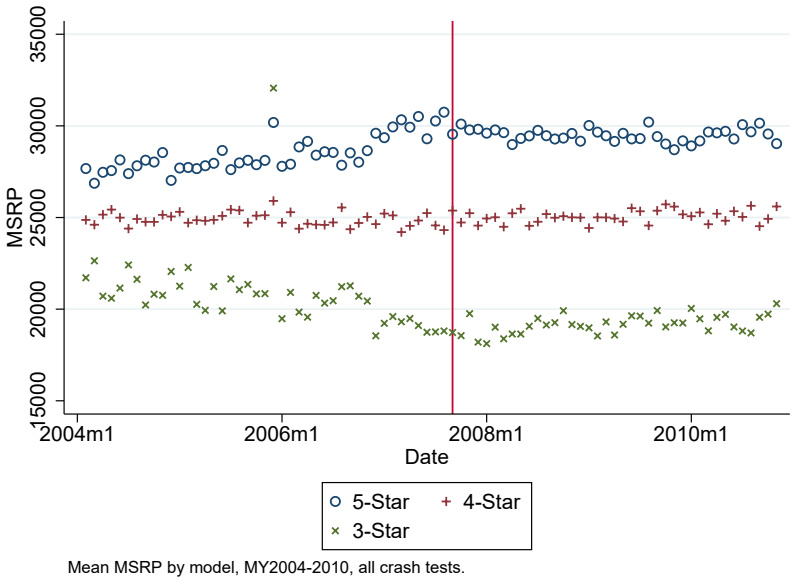


Figure 2.5: MSRP Continuity Check

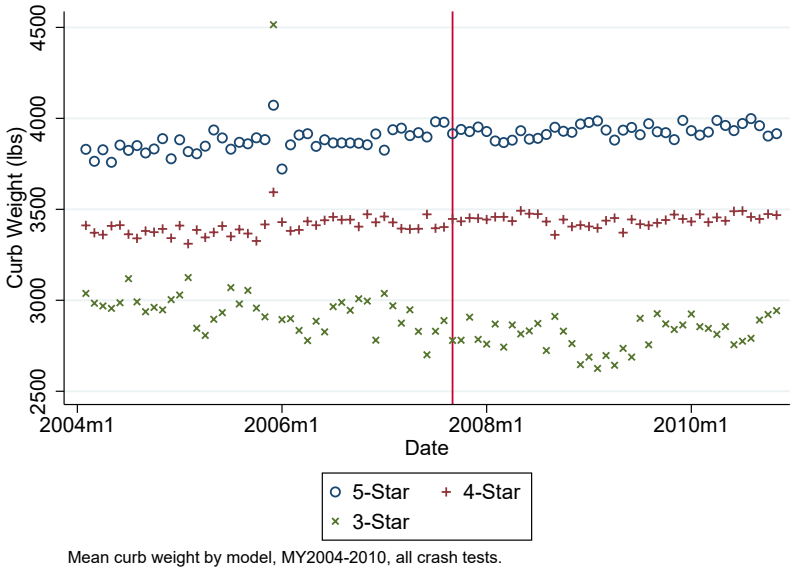


Figure 2.6: Curb Weight Continuity Check

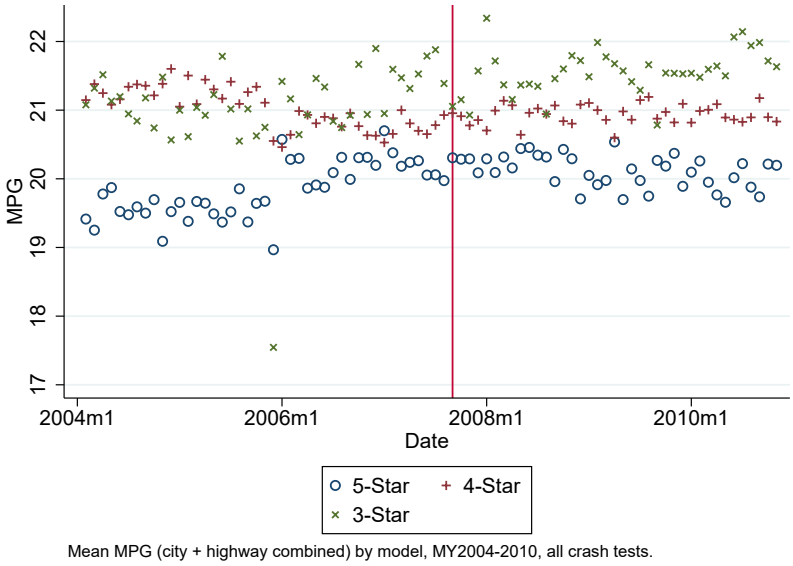


Figure 2.7: MPG Continuity Check

My outcome of interest is vehicle sales. The continuity assumption requires that covarites be continuous across the underlying running variable threshold in order to accurately calculate estiamtes via regression discontinuity. However, any sales response may be biased by changes in other vehicle characteristics. First, therefore, I examine whether NCAP affects MSRP, curb weight, fuel efficiency, and horsepower.

Producers may simply offset a negative star rating by dropping price, dampening any observed sales effect. Figure 2.8 plots mean MSRP by normalized probability of injury (the underlying, continuous running measurement from the NCAP crash test that is then mapped to a 0 5-star safety rating). I predict a negative correlation between probability of injury and MSRP. This is reflected in the data. I find evidence of discontinuity in MSRP at the ratings threshold.

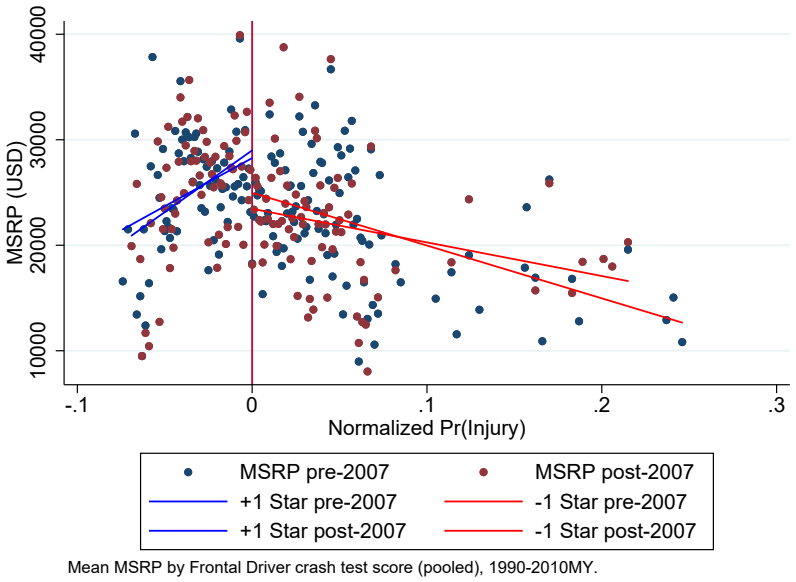


Figure 2.8: MSRP

Figure 2.9 plots mean curb weight against NCAP probability of injury. All else equal, I predict a negative correlation between curb weight and probability of injury. Producers may simply improve vehicle safety by increasing vehicle weight; if manufacturers also invest in fuel-saving technology to keep fuel efficiency high despite weight vehicle weight increases. [48] I find some evidence of discontinuity in curb weight at the ratings threshold. I am unable to conclude that manufactuers are not responding to NCAP by simply increasing curb weight.

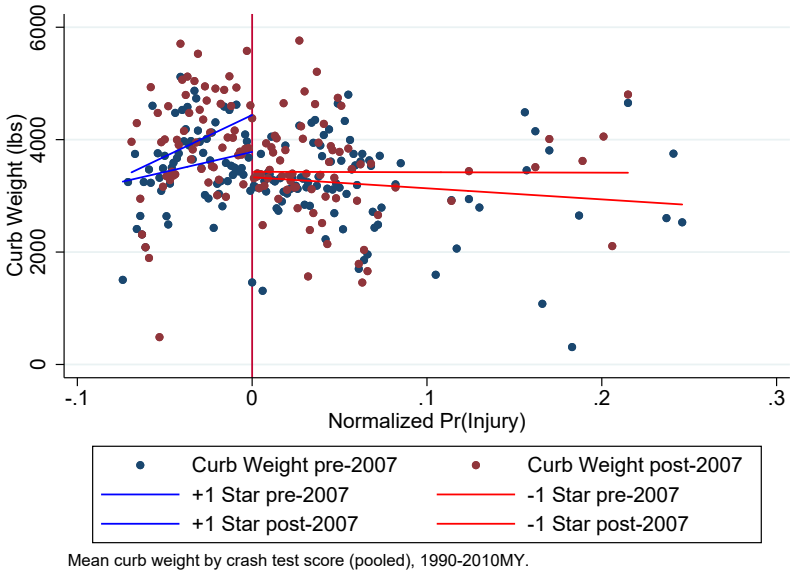


Figure 2.9: Curb Weight

Figure 2.5 plots mean fuel efficiency (miles per gallon, city and highway driving combined) against NCAP probability of injury.

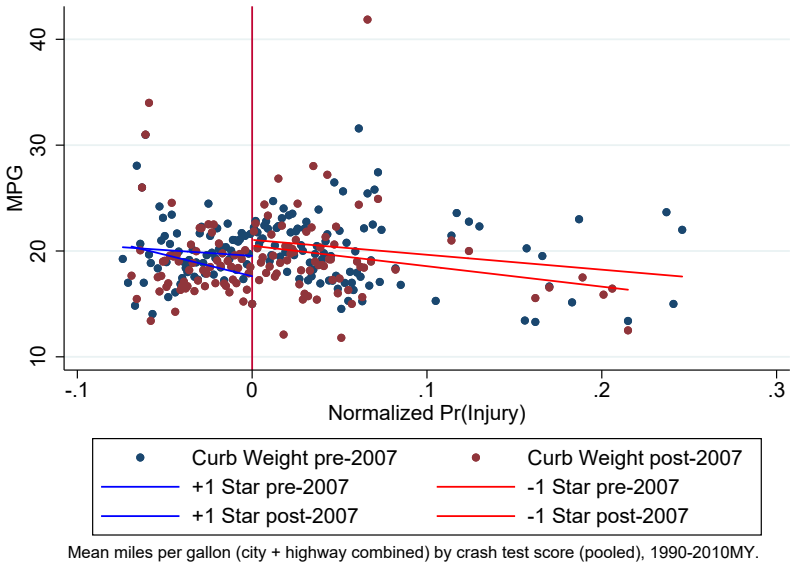


Figure 3.5.1 plots mean maximum horsepower against NCAP probability of injury.

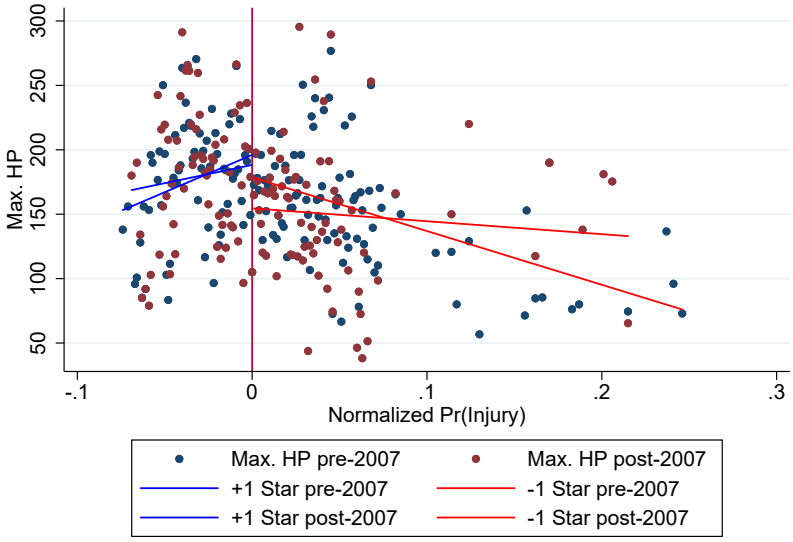


Figure 2.10: Horsepower

Figure 2.1 provides my main results. I am unable to conclude that a favorable NCAP rating has a significant positive effect on model sales.

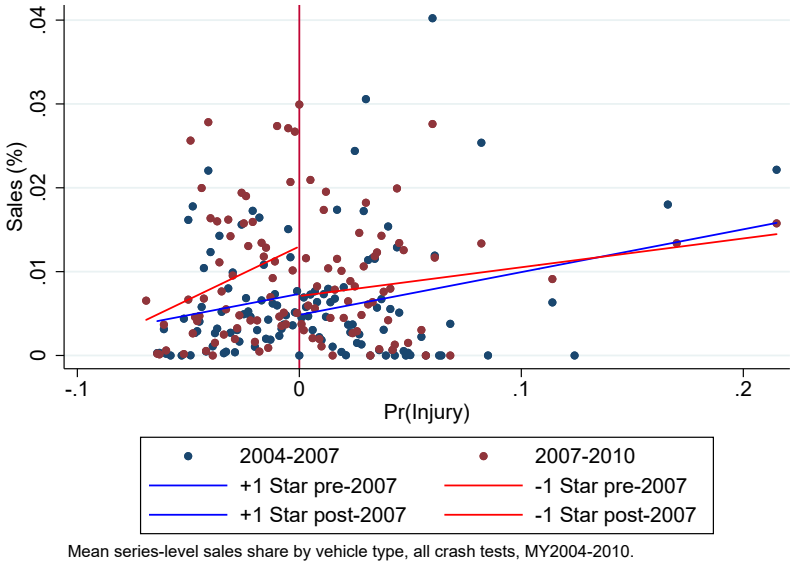


Figure 2.11: Results: Sales RD

Table 2.1 provides my key results for sales response by NCAP crash test type. I do not find a consistent, significant effect on sales of the scoring an additional star on the six NCAP crash test types. This may be due to manufacturers passing safety improvements through to the consumer via price, downwardly biasing any estimates of the effect of NCAP safety rating on sales.

Table 2.2 show the RD estimates in NCAP Pr(Injury) and star rating for curb weight, maximum horsepower, fuel efficiency, and MSRP. Only curb weight varies significantly with Pr(Injury); there is a strong quadratic relationship. This finding suggests that vehicle manufacturers are responding to the NCAP ratings by increasing curb weight.

Finally, figure 2.3 provides the RD estimates for sales, MSRP and curb weight

VARIABLES	(1) FD	(2) FP	(3) SD	(4) SP	(5) Roll	(6) Roll 4WD
Pr(Injury)	0.953	-0.695	0.145	0.613	-0.472	-0.886
	(0.649)	(0.464)	(0.252)	(0.684)	(0.495)	(5.540)
Pr(Injury)_sq	-1.828	2.193**	-0.570	-2.384	0.970	-11.64
	(1.957)	(1.065)	(0.922)	(2.581)	(1.074)	(90.07)
+1 Star	-0.0195	0.0975	0.0254	0.0197	-0.00109	-0.0773
	(0.0158)	(0.0788)	(0.0188)	(0.0189)	(0.0391)	(0.105)
+1 Star * Pr(Injury)	-1.372	6.145	-2.524	-1.417	-0.802	-2.650
	(1.060)	(4.355)	(2.907)	(1.564)	(2.471)	(5.038)
+1 Star * Pr(Injury)_sq	-1.358	50.88	-66.65	-9.672	-35.36	-13.85
	(16.21)	(56.08)	(73.78)	(28.92)	(42.25)	(152.4)
Constant	0.0513***	0.0589***	0.0253***	0.0138***	0.0402***	0.0895
	(0.0141)	(0.0214)	(0.00640)	(0.00470)	(0.0121)	(0.0759)
Observations	32,604	37,346	15,764	13,271	13,841	9,477
R-squared	0.090	0.134	0.053	0.048	0.019	0.065

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.1: Sales RD

VARIABLES	(1) Curb Weight	(2) Max HP	(3) MPG	(4) MSRP
Pr(Injury)	-12,835**	-510.5	-17.16	-43,660
	(5,074)	(427.6)	(18.60)	(45,060)
Pr(Injury)_sq	29,959**	993.7	19.60	74,154
	(12,039)	(1,141)	(44.20)	(123,138)
+1 Star	70.05	-9.715	-1.477*	-158.3
	(160.7)	(16.65)	(0.881)	(2,085)
+1 Star * Pr(Injury)	6,089	606.7	24.64	-58,305
	(17,415)	(1,485)	(69.56)	(156,194)
+1 Star * Pr(Injury)_sq	-122,901	6,085	328.2	-1.328e+06
	(293,686)	(26,561)	(1,304)	(3.020e+06)
Constant	3,452***	187.8***	21.31***	24,174***
	(144.0)	(11.30)	(0.804)	(1,292)
Observations	122,272	122,303	110,754	122,272
R-squared	0.073	0.010	0.018	0.021

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.2: RD - Vehicle Attributes

for both pre- and post- September 2007. Interestingly, both curb weight and MSRP show a large, significant negative response to NCAP Pr(Injury) in the post period. This provides evidence that consumers are responding to the Monroney sticker 2007 mandate.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Curb Weight	Curb Weight (post)	MSRP (post)	MSRP	Sales (post)	Sales (post)
Pr(Injury)	-5,876	-21,575***	-1,542	-96,027**	0.585	0.286
	(5,858)	(6,538)	(57,988)	(37,012)	(0.499)	(0.427)
Pr(Injury)_sq	10,657	52,871***	-22,420	200,893*	-1.241	0.0126
	(13,150)	(16,147)	(157,835)	(103,360)	(1.349)	(1.225)
+1 Star	-136.7	334.8	-2,106	2,559	0.0462	0.00588
	(178.6)	(226.5)	(2,472)	(2,450)	(0.0393)	(0.0115)
+1 Star * Pr(Injury)	-33,614**	59,184**	-343,288*	309,921	0.293	-0.520
	(16,713)	(28,610)	(202,288)	(192,803)	(2.068)	(0.864)
+1 Star * Pr(Injury)_sq	-632,994**	583,534	-5.050e+06	3.424e+06	-0.230	-10.58
	(289,110)	(452,703)	(3.921e+06)	(3.254e+06)	(37.40)	(10.60)
Constant	3,267***	3,719***	23,453***	25,002***	0.0479***	0.0193***
	(134.7)	(256.0)	(1,624)	(1,430)	(0.0124)	(0.00611)
Observations	52,746	69,526	52,746	69,526	52,766	69,537
R-squared	0.073	0.129	0.026	0.051	0.022	0.056

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2.3: RD + Event Study

2.6 Conclusion

This paper evaluates whether consumers respond to information about vehicle safety in making vehicle purchasing decisions. I evaluate this using (1) an event study around the 2007 Monroney Sticker mandate and (2) using an RDD around the NCAP star thresholds. I consider the sales response, any changes in price that might be dampening this estimated response, and technology adoption by manufacturers. I find no sales response at the NCAP ratings threshold. Rather, I find discontinuity in curb weight and MSRP at the star threshold, suggesting that manufacturers are increasing vehicle curb weight in order to perform better on crash tests and are passing the cost of "safety" improvements through to consumers via increased price.

In theory, greater transparency over product information leads to higher product quality. If this is true, and if NCAP does in fact increase transparency over vehicle safety, it would suggest a role for regulatory intervention in correcting product market information asymmetries as a channel by which to promote technological innovation. However, the relationship between product quality, innovation and competition is ambiguous.

Can public policy incentivize the adoption of life-saving vehicle safety technology, and if so, what is the most cost-effective way to achieve this? My findings suggest a role for regulatory intervention in standard-setting, and also in correcting for information asymmetries as a channel to promoting technological innovation.This paper also highlights the importance of designing policy interventions that target technological

development to take into consideration where along the technology adoption life-cycle a certain technology, product or market might be. For example, what is the optimal testing regime and rating system for the NCAP program? For example, a longer testing timeframe would increase manufacturers' incentive to score above the star threshold, which may strengthen supply-side response. Further, if NCAP generates a negative externality born disproportionately by drivers of older, less-safe vehicles, does this suggest a government role in subsidizing vehicle safety for less-safe vehicles?

It may be more effective, for example, to set and enforce crashworthiness standards than to subsidize emergency response costs. Is it more cost effective to invest in programs to promote private-sector innovation in safety technology, or invest in infrastructure, law enforcement, etc.? If the answer is dependent on stage of technological innovation and adoption of various crash prevention and crashworthiness technologies, then there is potential for policy reform to more accurately promote vehicle safety.

There are significant equity/distributional consequences to regulation of safety technology in U.S. passenger vehicles. It is the role of government to intervene in product markets in order to ensure public safety. Fortunately, economics provides a number of channels for this intervention, from programs and standards designed to decrease information asymmetry, to policies that incentivize innovation in health and safety. Federal safety regulation has played an active role in the U.S. automobile market since its inception, from speed limits, to seatbelt laws, to airbag and other federal motor vehicle safety standards. Further, production constraints can lead to a tradeoff between vehicle safety and fuel efficiency. As a result, fuel economy standards can also have an indirect

effect on vehicle safety. Programs such as NCAP that establish a federal mandate and avenue for the disclosure of vehicle safety information can alleviate information asymmetry, but may have adverse effects, such as causing producers to increase curb weight as a low-cost way to increase vehicle crashworthiness (citation). In this case, there is a negative externality to a "vehicle arms race" to improve crashworthiness as the return to safety is positively correlated with the number of heavier vehicles (e.g. SUVs) on the roads.

Is there a role for government intervention to promote vehicle safety, or is information sharing in the status quo sufficienct to promote an optimal level of innovation in the auto market? Is it more effective to target the consumer or producer side of the market in designing policy to promote safer automobiles? Safety information can provide a positive externality, suggesting a role for programs such as NCAP that standardize/facilitate/subsidize accurate information sharing with consumers. Further, given the extended production timeframe in the automobile market, there are distributional and welfare effects as firm response (in terms of safety technology adoption) is seen in the discontinuing or introduction of models, and trickles down to the used car fleet and onto public roads. Given the U.S's position as a leader in the auto industry, federal vehicle safety regulations such as NCAP have implications beyond U.S. markets, seatbelts, helmets in other countries.

One possible outcome of NCAP is that producers may increase curb weight when they receive and unfavorable rating in order to improve crashworthiness, the safety outcome that NCAP measures. This could potentially lead to a vehicle weight "arms

race" with negative externalities in terms of safety (for other vehicles) and automobile emissions [43] Therefore any attribute-based policy must consider general equilibrium effects throughout the industry, not just in terms of vehicle safety.

The U.S. is a market leader in automobile industry standards, and the NCAP program has significant implications beyond the U.S. for international standardization of vehicle safety evaluation regimes. Versions of NCAP exist as Euro NCAP, China NCAP, Japan NCAP, Korean NCAP, Australasian NCAP, Latin America and Caribbean NCAP.

I am also interested in whether the NCAP safety ratings affect manufacturers' adoption of life-saving technologies. Does NCAP affect the diffusion of safety technologies throughout the fleet of new vehicles (e.g. from high-end to low-end vehicles)? Do firms shift to crash prevention technology, which may be more cost-effective than continued investment in crashworthiness technology, given different S-curves of technology adoption? Manufacturer response likely depends on appropriability and imitation costs of safety and other vehicle technology.

Perhaps a good place to start is by asking whether NCAP accurately signals information about vehicle safety. In Chapter 3, I use data from NHTSA's Fatal Accident Reporting System (FARS) and the Texas Department of Transportation's Crash Records Information System (CRIS) to evaluate the correlation between NCAP ratings and real-world crash outcomes. FARS is a national reporting system for all vehicle collisions in the U.S. involving a fatality, while CRIS has fatal and non-fatal vehicle collision data for Texas. [45]

Chapter 3

Test Driving the New Car Assessment Program

3.1 Introduction

Automobile accidents are the leading cause of death via unintentional injury in the U.S. [22] In 2016, there were 37,461 motor vehicle fatalities in the U.S., or 1.18 per 100 million vehicle miles traveled (VMT). This represented a 5.6% increase from 2015 relative to a 2.2% increase in VMT. In 2015, there were 35,092 fatalities in 2015 (1.12 per 100 million VMT), and 2,443,000 injuries in automobile accidents. The US Center for Disease Control (CDC) estimates that the 33,687 fatalities due to motor vehicle accidents in 2010 incurred a medical cost of over $374 million. For non-fatal hospitalized injuries in 2010 for motor vehicle occupants, of which there were 178,505, the CDC estimates an average medical cost of $54,197. For nonfatal Emergency

Department treated and released injuries for motor vehicle occupants in 2010, of which there were 2,557,616, the CDC calculates an average medical cost of $3,222. [22]

Given the increasing importance of transportation to public health, the transportation industry is an important yet often overlooked consideration in today's economic and political climate, particularly in its implications for public health and socioeconomic equality. Federal safety regulation has played an active role in the U.S. automobile market since its inception, from speed limits, to seatbelt laws, to airbag and other federal motor vehicle safety standards. As highlighted in Chapter 2, government intervention in the market for vehicle safety can help correct for information asymmetries between producers and consumers. There are thus significant public health and economic consequences to automobile safety standards and regulation. Safety ratings can provide a demand-side approach to improving safety. Safety is primary factor in consumer choice over vehicle purchase [49], [41], [80], and safety information can provide a positive externality for public health and spending. Aggregated and standardized ratings decrease the cost of information acquisition and transfer.

Vehicle safety technology has significant and widescale implications for public health and regulatory policy. U.S. regulators can influence vehicle safety via two mechanisms: (1) safety standards (such as seatbelts, airbags or safety ratings); and (2) recalls or bans of dangerous vehicles. In keeping with (1) and as discussed in Chapter 2, the U.S. National Highway and Transportation Administration (NHTSA) evaluates the safety of all new vehicles sold in the United States. NHTSA publishes these safety ratings for new vehicle models sold in the U.S. on a 5-Star scale for frontal and side

driver and passenger, and rollover, crash tests via the New Car Assessment Program (NCAP). The 5-Star NCAP vehicle safety ratings are published on the NHTSA website, in numerous consumer reports, and are required to be displayed in the driver's window of all new vehicles sold in the U.S.

In this paper, I match NCAP safety ratings with real-world crash data from the Fatality Analysis Reporting System (FARS) and Texas' Crash Records Information System (CRIS) in order to determine whether there is a correlation between NCAP rating and vehicle-level crash outcomes, and whether NCAP rating is an accurate predictor of risk of injury and/or death for vehicle occupants. The FARS database represents all car crashes in the U.S. involving at least one fatality, while the Texas CRIS data represents all police-reported car crashes in Texas.

A significant limitation to this approach is that if the NCAP program changes consumer behavior, as addressed in Chapter 2, this will bias any estimates of the true relationship between NCAP Pr(Injury) and real-world crash outcomes. For example, if safer drivers are willing to pay a premium for safer vehicles, and if these consumers are more likely to be influenced by the NCAP program, then any finding of a relationship between NCAP rating and real-world crash outcomes may over-state the true correlation between predicted risk of injury and real-world injury, fatality and vehicle damage due to driver selection.s Similarly, if more dangerous drivers tend to purchase certain vehicle models and drive them more recklessly, these models may be falsely deemed less "safe" due to driver selection. Although future work will aim to incorporate vehicle registration data, in this paper I am unable to control for driver behavior due to data limitations.

Further, if FARS and CIRS crash reports do not include less severe crashes with no fatalities (FARS) or no injuries/less than $1000 in property damage (CIRS), then data under-represents less severe crash outcomes, which in turn are more likely to involve safer vehicles.

Nevertheless, a strong comparison of NCAP crash test results to real-world crash outcomes seems warranted. After all, how valuable are NCAP ratings to consumers if they are based on highly restrictive and simulated crash conditions which bare little resemblance to real-world accidents? [60] I find minimal correlation between NCAP Pr(Injury) and occupant injury and vehicle damage outcomes. I find a strong negative correlation between curb weight and injury and damage rates. Further, curb weight is significantly negatively correlated with Pr(Injury), as well as fuel efficiency. Drawing on my results from Chatper 2, I am left to conclude that manufacturers respond to NCAP by increasing vehicle weight as a way to improve model crash test scores.

The remainder of this Chapter is organized as follows: Section 2 reviews the literature and discusses my contribution. Section 3 discusses NCAP and FARS. Section 4 introduces the data and outlines my empirical strategy. Section 5 presents my results. Section 6 concludes.

3.2 Previous Literature

I examine whether the New Car Assessment Program's safety ratings accurately predict real-world crash outcomes in terms of vehicular damage and occupant injury and fatality rates. As discussed in Chapter 2, NCAP's 5-star vehicle safety ratings are highly salient so consumers. They are federally mandated to be displayed alongside mileage on the Monroney window sticker of each new car sold in the U.S., and are widely published in consumer reports and advertisements. Consumers use NCAP ratings in making the new vehicle purchasing decision, and producers design new vehicles to perform well on NCAP crash tests. However, these crash tests are conducted in a highly simulated laboratory setting hat may not reflect real-world conditions.

Crash testing is conducted under highly simulated laboratory conditions. Further, data on crash outcomes from police reports lack detailed information on occupant injury, and may over-classify injuries as "serious" that under the Abbreviated Injury Scale (AIS) would generally be considered "minor". [60] Crash reports do not report body region of injury; however, crash test ratings are calculated based on measurements taken from crash dummies, which indicate risk of injury to specific regions of the body. Unsurprisingly, previous studies have found discrepancies between crash ratings and real-world crash outcomes.

An innovative safety technology is often implemented at first by a single manufacturer or as a standard feature for luxury models. Only with enough data to provide evidence that the technology is actually effective in reducing injuries or deaths does

NHTSA enter the extended process of mandating a given technology's includsion in all future model-year vehicles. This process invovles federally mandated rule-making procedures and a staggered implementation and enforcement period that is coordinated with the automobile industry's cyclical design timeframe. The time for new safety technology to "trickle down" to penetrate the used vehicle fleet is even longer. [38] According to NHTSA, it takes a median of 12.5 years for cars to age out of circulation in the United States [**?**].

Previous studies have shown a lower risk of injury and death for higher-rated vehicles. (Campbell 1982, Jones & Whitfield 1988, Lie & Tingvall 2002, Kullgren et al. 2010, Ryb et al. 2010, Metzger et al. 2015, Farmer 2006) [19] Jones & Whitfield (1988) compare NCAP risk of injury to driver to 6,405 crashes in Texas, controlling for vehicle mass, driver age, restraint use and accident severity. [45] NCAP rating was calculated based on Head Injury Criterion (HIC), Chest Deceleration (CD) and femur load to crash dummies. However, the authors restrict their analysis to estimating the risk of incapacitating injury or death for drivers in single-car, fixed-object, rontal collisions. A companion paper finds similar results using a sample of fatal frontal collisions between two vehicles of similar mass.

Lie & Tingvall (2002) evaluate the accuracy of the Euro NCAP in predicting injury and death outcomes. They consider severe and fatal injury outcomes to driver and front-seat passenger in two-car crashes (they find no correlation between rating and minor injury outcomes). Using a sample of 64 Euro NCAP-rated vehi cle models. They find that occupants of vehicles rated 3 or 4 stars (at the time of the study Euro

NCAP has a maximum of 4 stars) are significantly less likely to experience severe or fatal injuries than vehicles rated 2 stars. However, 1200 of approx. 1500 cases in their sample involve vehicle that had not been rated. Further, the authors do not estimate results separately by vehicle class (curb weight). [54]

Metzger et al. (2015) evaluate the correlation between NCAP rating and real-world risk of injury to rear seat occupants. Using a sample of 18,218 vehicles NASS-CDS (a nationally representative sample of approximately 5,000 police-reported vehicle crashes annually involving at least one towed vehicle and causing property damage and/or injuries. NASS over-samples certain types of crashes, including crashes involving fatalities and/or hospitalizations. They restrict their sample to 5- vs. 4-star vehicles and to restrained vehicle occupants. They find a slightly lower probability of injury for rear-seat occupants of vehicles that received a 5-star rating from the frontal driver and passenger impact tests (relative to 4-star). They find no correlation between side driver and passenger NCAP rating (the only crash test that incorporated measurements from a rear-seat crash dummy) and risk of injury to a rear-seat passenger. They conclude that updating the NCAP rating system to better reflect rear seat risk of injury should be prioritized. [57]

Ryb et al. (2010) compare Insurance Institute for Highway Safety (IIHS) crash test ratings to mortality rates in real-world crashes. They look at injury to drivers in frontal crashes using a detailed sample from the Crash Injury Research Engineering Network (CIREN). IIHS rates vehicles as "good", acceptable, marginal and poor. They find that vehicles rated good or average experienced lower mortality rates. However,

over half of their sample of 1,276 drivers were driving non-rated vehicles.

Matching vehicle safety ratings from the Insurance Institute for Highway Safety, FARS, and vehicle registration data, Farmer (2006) find significantly lower fatality rates for "poor" vs. "good" rated vehicles. [34]

Kullgren et al. (2010) use discrete 5-star Euro NCAP score (not continuous probability of injury) to compare relative risk of injury calculated from crash reports and insurance claim data. They find a lower risk of injury and mortality for higher Euro NCAP-rated vehicles. [51]

Newstead et al. (2003) consider both NCAP and IIHS frontal crash test ratings and real-world accident outcomes in Florida, Ohio and Pennsylvania. They do not find a significant correlation between crash test performance and real-world injuries. The authors argue that this disparity is due to the lack of detailed information on occupant injury in real-world police-reported crashes. [60] However, this disparity results in part from the highly simulated conditions of the crash testing relative to real-world driving behavior.

Injury severity depends on a number of factors besides vehicle crashworthiness. Vehicle speed and configuration at the time of the crash, vehicle size and weight, and restraint use by vehicle occupants may have a greater effect on injury probability and severity than just crashworthiness. [60]

Lack of correlation between NCAP predicted probability of injury and real-world crash outcomes could suggest the need for reform of the NCAP program. Krafft et al. (2000) propose an alternative method to deriving risk functions based on change

in vehicle velocity during crash. [50]

3.3 The New Car Assessment Program

NHTSA's New Car Assessment Program provides comparative information on the safety of new vehicles to assist consumers with vehicle purchasing decisions and encourage motor vehicle manufacturers to make vehicle safety improvements. Its mission is to: (1) Help consumers with vehicle purchasing decisions, (2) Incentivize manufacturers to improve current safety performance and features of new vehicles; and (3) Promote innovation and development of new vehicle safety features. [6]

NCAP was established 1978 under Title II of the Motor Vehicle Information and Cost Savings Act of 1972. NHTSA began testing vehicles for crashworthiness using frontal driver and frontal passenger crash tests with 1979 model year. 5-Star ratings were introduced in 1994 for MY1990- vehicles.

A variety of ratings systems exist, both in the US and internationally. There are both private and government-regulated ratings, for new and used vehicles, from both experts and consumers. NCAP is a government-run program, but industry ratings are also published. Edmunds provides both new and used car ratings, including peer/owner ratings out of 5 stars by 0.5-star increments. While NCAP is a predictive rating system, there are also retrospective ratings based on crash reports and insurance claims. For example, NHTSA's Office of Defects Investigation (DOI) has an Early Warning Reporting (EWR) to monitor fleet performance post-sales. NCAP focuses on crashworthiness, but other ratings programs evaluate crash avoidance and prevention technologies.

The primary rating program other than NCAP in the US is the Insurance

Institute for Highway Safety (IIHS), which began publishing a frontal offset impact rating in 1995. In addition to evaluating crash worthiness like NCAP, IIHS also rates vehicles for crash avoidance and mitigation. In terms of crashworthiness, IIHS rates vehicles Good, Acceptable, Marginal or Poor based on performance on six crash tests: driver-side small overlap front, passenger-side small overlap front, moderate overlap front, side, roof strength and head restraints. In terms of crash avoidance, IIHS rates all vehicles with available front crash prevention systems either Basic, Advanced or Superior, based on the system type and vehicle performance in crash tests. IIHS also tests and rates headlights as Good, Acceptable, Marginal or Poor.

Due to the complexity of vehicle safety technology and other vehicle characteristics, transaction and search costs for individual consumers in the new vehicle purchasing decision can be high. Ratings program such as IIHS can be complex. A primary goal of the 5-star NCAP ratings was to simplify and standardize the information about vehicle safety that consumers are exposed to when purchasing a new vehicle.

NCAP focuses on crashworthiness technology such as seatbelts and airbags, but also identifies whether rated vehicles are equipped with Crash Avoidance Technologies, such as Electronic Stability Control (ESC), Lane Departure Warning (LDW), and Forward Collision Warning (FCW). Additional crash avoidance technologies include rearview video systems (RVS), automatic emergency breaking, and tire pressure monitoring systems (TPMS).

NCAP calculates safety ratings for each vehicle based on three separate crash tests: frontal, side and rollover. The frontal driver and passenger crash test involves

crashing the tested vehicle into a fixed crash barrier and is conducted at 35mph (56.3 km/h). The driver and front-seat passengers are average-sized adult male dummies and rating is an evaluation of injury to the head and chest (discussed further below). Only vehicles from the same weight class can be compared. The frontal test simulates a head-on collision between two similar vehicles each traveling at the same speed.

The side driver and side passenger crash tests were added for MY1997. In the Side Impact crash test, all vehicles are hit with the same force by a moving barrier or pole, so rating results can be compared across all classes. The crash test is conducted at 38.5mph and is intended to simulate an intersection-type crash. The driver is an average-sized adult male dummy and the passenger is an average-sized adult male dummy in the rear seat, and rating is calculated based on measurements on injury to the chest. This is the only crash rating that measures injury to a rear-seat occupant.

NCAP introduced rollover ratings for MY2001- vehicles. The Rollover crash test assesses risk of rollover and is calculated based measurements taken from driver and front-seat passenger dummies. Rollover ratings can be compared across vehicle classes.

As of September 2007, NCAP 5-star ratings are required to be displayed in the driver's side window of all new vehicles sold in the US. In 2011, NHTSA overhauled the NCAP ratings formulae and introduced a new "Overall" star rating for MY2010- vehicles as a weighted average of the front, side and rollover ratings. In this paper, I restrict my analysis to MY 1994-2010 vehicles and therefore do not use the overall rating category.

For each of the three crash test types, measurements taken from driver and

front passenger dummies are used to generate six types of crash test rating types: Frontal Driver, Frontal Passenger, Side Driver, Side Passenger, Rollover and Rollover 4WD. In 2011, and Overall rating was added as a weighted sum.

Risk of injury on each Test is calculated for Front and Rear Driver and Passenger(s). Instruments measure the force of impact to each crash dummys head, neck, chest, pelvis, femur (legs), and feet:

$$P(Injury) = f(HIC, TTI, CD, PD, FL)$$

where HIC is Head Injury Criterion, TTI is Thoracic Trauma Index, CD is Chest Deceleration, PD is Pelvis Deceleration, and FL is Femur Load.

The Probabilities of Injury on each Test type T = Front, Side, Rollover are calculated as the average of the Driver and Passenger crash performance scores:

$$P_T(Injury) = 0.5P_{T,Driver}(Injury) + 0.5P_{T,Passenger}(Injury)$$

For the Frontal Crash Test:

$$P_{Driver}(Injury) = (1 - (1 - P_{Head}(Injury)) * (1 - P_{Chest}(Injury)))$$

where $P_{Head}(Injury)$ and $P_{Chest}(Injury)$ are calculated for both Driver and

Passenger according to the formulae:

$$P_{Head}(Injury) = 1/(1 + 1/(1 + exp(5.02 - 0.00351HIC_{Front})))$$

$$P_{Chest}(Injury) = 1/(1 + 1/(1 + exp(5.55 - 0.0693CD_{Front})))$$

Post-2010, Overall vehicle performance is calculated from the front, side and rollover crash tests according to the formula:

$$P_{Overall}(Injury) = 0.42P_{Front}(Inj.) + 0.33P_{Side}(Inj.) + 0.25P_{Roll}(Inj.)$$

3.4 Data & Methodology

I construct a novel dataset using underlying probability of injury, the continuous running variable used to calculate the NCAP 5-star ratings from vehicle crash test performance. I match NCAP probability of injury with two databases of police crash reports: (1) NHTSA's Fatality Analysis Reporting System (FARS); and (2) Texas Department of Transportation's Crash Records Information System (CRIS).

I get data on crash outcomes from the National Highway Transportation and Safety Administration (NHTSA)'s Fatality Analysis Reporting System (FARS). FARS provies detailed police crash reports on all motor vehicle accidents occuring on public roads in the U.S. involving at least one fatality, 1975-2018. FARS accident variables include manner of collision, type of intersection, light condition, atmospheric conditions, speed limit, roadway condition, and other accident-level variables. FARS vehicle-level variables include vehicle travel speed at the time of the crash, areas and extent of damage, etc. Manner of collision in the FARS dataset is dependent on the directions of travel of the vehicles involved and the geometry of the points of impact, and is divided into Rear-end, Head-on, Rear-to-Rear, Angle, Sideswipe and same vs. opposite-direction.

They key vehicle-level outcome variables I use in my analysis are vehicle damage and number of deaths. Vehicle damage in the FARS data is coded as none, minor, functional or disabling.

I calculate vehicle damage outcomes as total counts over total number of

crashes by model within a 2-year window of vehicle model year:

$$Y_{i,MY} = \sum_{t=MY}^{MY+2} (y/Crashes)_{it,MY}$$

where $Y_{i,MY}$ is my outcome of interest, vehicle damage and death rates by model i and model year MY; $Crashes_{it,MY}$ is the total number of crashes by vehicle model and crash year t, inclusive; and $y_{it,MY}$ is total number of deaths and counts of vehicle damage for model i and model year MY in crash year t.

FARS Person-level variables on driver and passengers include age, gender, height, weight, seating position, restraint use, and injury severity. There is also detailed information on driver history, including previous motor vehicle citations. Crash tests are done with belted crash dummies. I present results for both restrained and unrestrained vehicle occupants.

The key person-level outcome variable I use in my analysis is injury severity, which is coded categorically as no injury, possible injury, incapacitating injury, non-incapacitating injury, or fatal injury. As with vehicle damage, I calculate occupant injury outcomes as total counts per occupant over total number of crashes by model within a 2-year window of vehicle model year:

$$Y_{i,MY} = \sum_{t=MY}^{MY+2} (y/(Crashes * Occupants))_{it,MY}$$

where $Y_{i,MY}$ is my outcome of interest, occupant injury rates by model i and model year MY; and $y_{it,MY}$ is total injury count for model i and model year MY in crash

year t. I calculate $Y_{i,MY}$ for no injury, possible injury, minor injury, non-incapacitating injury, incapacitating injury, and fatal injury counts.

The FARS data is not representative of all crashes. It under-represents less severe crashes, as it only includes crashes in which at least one person was killed. To overcome this reporting bias, I use data from Texas' Department of Transportation's Crash Records Information System, a database of all motor vehicle collisions, not just accidents resulting in fatalitiies.

Texas' DOT CRIS contains all data collected from the Texas Peace Officer's Crash Report (CR-3). Texas Transportation Code 550.062 requires any law enforcement officer who in the regular course of duty investigates a motor vehicle crash that results in injury to or death of a person or damage to property to the apparent extent of $1000 or more to submit a written report of that crash to the TxDOT within 10 days of the crash. Form CR-2 is the equivalent form required of the driver of a motor vehicle in a meeting the above conditions not investigated by a law enforcement officer. The TX CRIS data thus under-represents minor crashes that involve no deaths or injuries and under $1000 in property damage. The key CRIS outcome variables I use in my analysis are injury rates and vehicle damage. Injury categories are unknown, no injury, possible, nonincapacitating, incapacitating, and death. Vehicle damage is coded on a scale of 1 = minimal damage to 7 = maximum damage.

As in Chapter 2, I get information on model-level vehicle characteristics and safety technology from DataOne, which provides vehicle characteristics at the VIN-prefix level for the universe of light duty vehicles produced 1981-2016MY. These include

information on vehicle body style, drive train, and production release. I get model-level sales data from WARDS Automotive, which provides sales and production data for the U.S. and North America at the subseries level for MY1980-2017. I use vehicle safety ratings from NHTSA's New Car Assessment Program. Ratings are provided at the vehicle make and model level.This includes crash test results and 5-star safety ratings for all rated vehicles from 1990-2016.- Crash test results and 5-Star Safety Ratings for all rated vehicles 1990-2016. Combining NCAP rating and underlying probability of injury with detailed information on driver, occupant and crash characteristics allows me to control for any systematic differences in restraint use, driver characteristics, crash type, or other covariates across models that could potentially bias any estimates.

Table 3.1 provides summary statistics for the FARS data. I observe a total of over 400 unique vehicle models across MY1994-2010. The mean NCAP star rating is 3.98, with a mean Pr(Injury) of 0.15 based on all test types. The mean number of crashes per model within a 2-year window of MY is 182. Over 60% of vehicle occupants are the driver, 20% front passenger and 20% rear passenger.

Table 3.2 provides summary statistics for the TX CRIS data. Because I only have CRIS outcomes for 2010 onwards, my results for the TX data are for MY2008-vehicles. Because the CRIS data includes minor accidents in addition to fatal accidents, it is not surprising that the mean NCAP star rating for the TX data is 4.46 with a mean Pr(Injury) of 0.10. The mean number of crashes per model, 1555, is an order of magnitude greater than for the FARS data. The mean fatality rate, .004, is approximately half that of the FARS data.

Table 3.1: FARS Summary statistics

Variable	Mean	Std. Dev.	N
Model Year	2000.371	4.181	212766
Crash Year	2001.401	4.21	212766
Curb Weight	3591.403	1217.557	212523
Height	63.204	10.085	212523
MPG (city)	17.087	3.937	197865
MPG (highway)	23.698	5.36	197865
Max HP	147.753	93.41	212766
MSRP	22213.88	7707.572	212364
NCAP Star Rating	3.978	0.946	212763
NCAP Pr(Injury)	0.15	0.1	212766
Total Crashes	181.721	213.479	212766
Driver	0.609	0.488	212766
Passenger (Front)	0.198	0.399	212766
Passenger (Rear)	0.141	0.348	212766
No Injury	0.003	0.012	206208
Possible Injury	0.001	0.007	206208
Non-Incapacitating Injury	0.002	0.009	206208
Incapacitating Injury	0.001	0.007	206208
Fatal Injury	0.004	0.015	206208
Minor Damage	0.075	0.263	212766
Functional Damage	0.143	0.35	212766
Disabling Damage	0.707	0.455	212766

Table 3.2: TX Summary Statistics

Variable	Mean	Std. Dev.	N
Model Year	2009.246	0.796	244328
Crash Year	2010.693	0.762	244328
Total Crashes	1555.143	1745.271	244328
Curb Weight	3885.828	1048.838	244327
Height	64.302	7.886	244327
MPG (city)	19.674	5.17	230592
MPG (highway)	26.964	5.787	230592
Max HP	220.852	75.209	244328
MSRP	26209.246	9073.752	244326
NCAP Star Rating	4.46	0.651	244328
NCAP Pr(Injury)	0.099	0.064	244328
Total Crashes	1555.143	1745.271	244328
Total Injuries	0.236	0.155	244328
Deaths	0.002	0.003	244328
Incapacitating	0.014	0.041	244328
Non-Incapacitating	0.059	0.033	244328
Possible	0.163	0.149	244328
No Injury	1.142	0.212	244328
Unknown	0.02	0.014	244328
Vehicle Damage (max=7)	2.246	2.299	239053

3.5 Analysis & Results

3.5.1 FARS

The FARS crash outcomes I consider are accident, fatality and injury rates (relative to both total US sales and total accidents by vehicle model), as well as vehicle damage (minor, functional, or disabling), as a function of NCAP Pr(Injury).

Figure 3.1 plots the distribution of NCAP Pr(Injury) for the FARS data. For the following FARS results, I restrict my analysis to $Pr(Injury) \leq 0.40$.

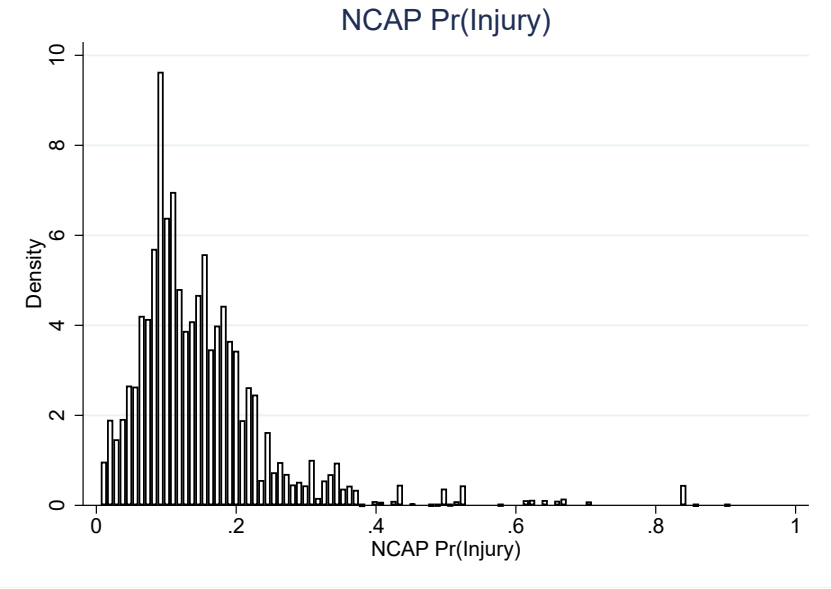


Figure 3.1: Distribution of NCAP Pr(Injury): FARS

It is important to note that NCAP Pr(Injury) is positively correlated with curb weight, as shown in figure 3.5.1. This is unsurprising, as heavier vehicles perform better on crash tests.

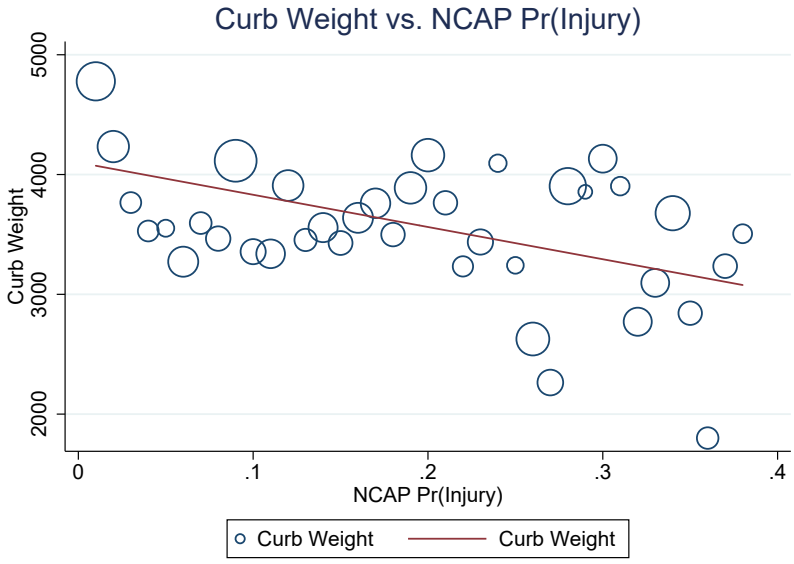


Figure 3.2: Curb Weight vs. NCAP Pr(Injury)

NCAP Pr(Injury) is negatively correlated with maximum horsepower, as shown in figure 3.5.1. This is opposite to what we would expect, as safety and performance are substitutes in manufacturer choice over vehicle attributes.

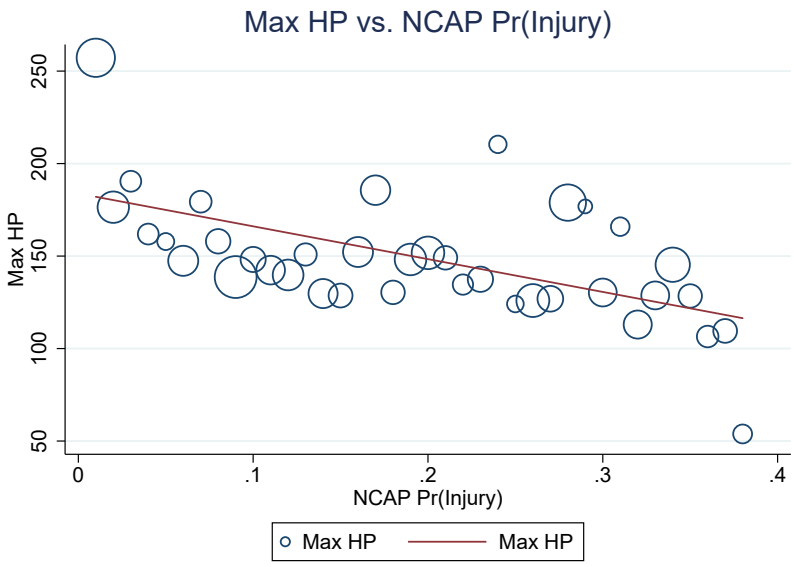


Figure 3.3: Max HP vs. NCAP Pr(Injury)

Figure 3.5.1 shows total crash count against NCAP Pr(Injury). Because NCAP measures crashworthiness, not crash avoidance, it is unsurprising to find no correlation between crash rate and NCAP Pr(Injury). However, this figure does not control for sales volume, which will skew crash counts towards more popular vehicle models.

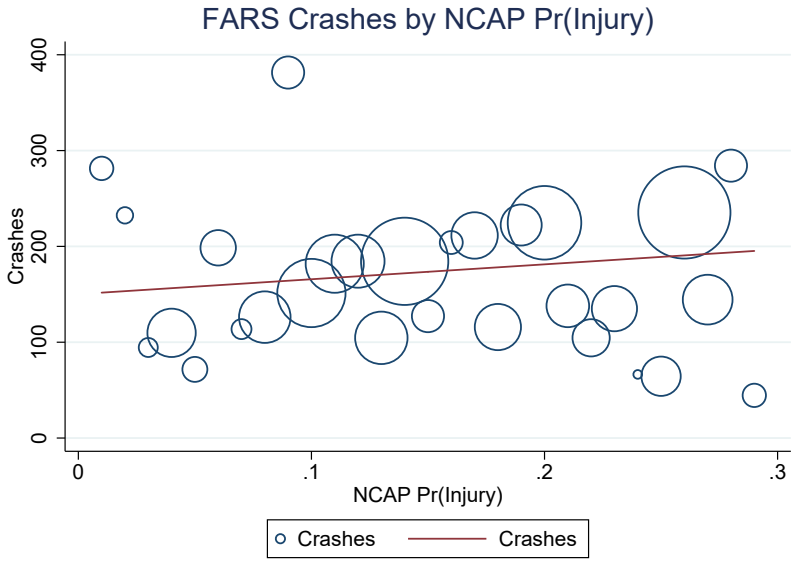


Figure 3.4: FARS Crash Count

As shown in Chapter 2, consumers may respond to NCAP safety ratings in purchasing, which is reflected in sales. If so, I would expect a jump in number of crashes at the ratings threshold of Pr(Injury). Figure 3.5 plots total sales against normalized, pooled NCAP Pr(Injury) for all test types and star levels by model. I find no evidence of discontinuity at the NCAP ratings threshold.

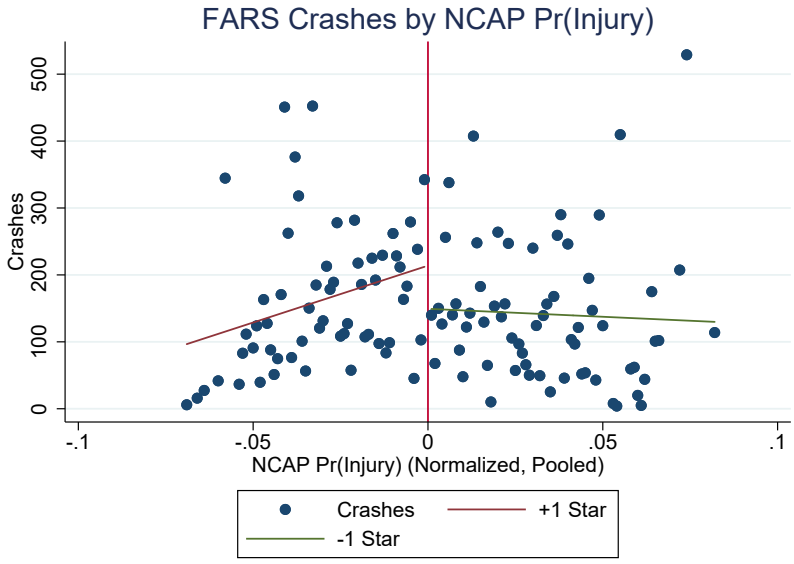


Figure 3.5: FARS Crashes: RD

Figure 3.6 plots FARS fatality rate against NCAP probability of injury, for all models, vehicle occupants and test types. There is a strong positive relationship between real-world fatality rate and NCAP predicted probability of injury. The quadratic relationship between fatality rate and Pr(Injury) appears to hold for low values of Pr(Injury).

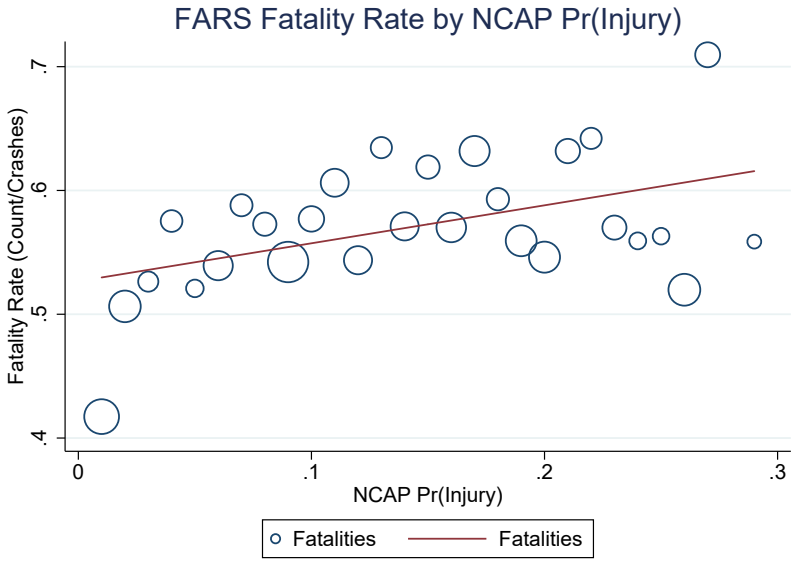


Figure 3.6: FARS Fatalities

Figure 3.7 provides the same results for disabling damage to vehicle. Disabling damage is increasing in Pr(Injury), as we would expect, as "disabling" is the most severe damage category.

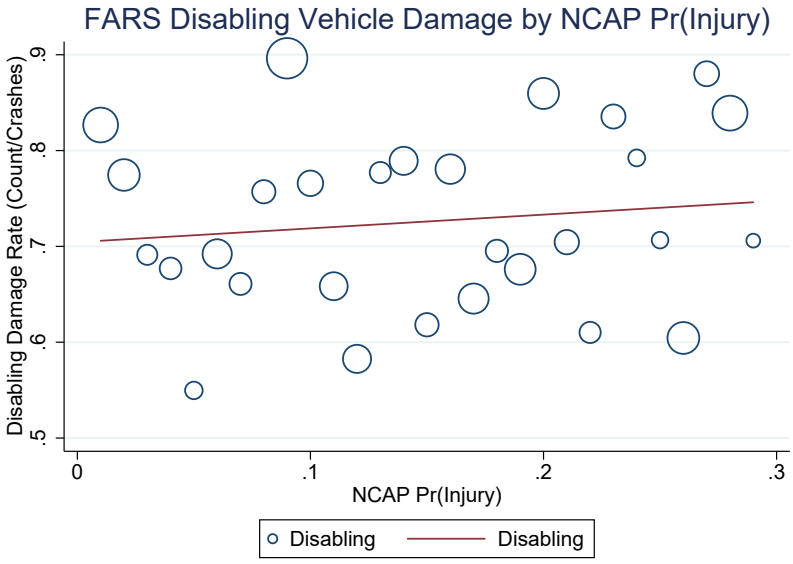


Figure 3.7: FARS Vehicle Damage: Disabling

Figure 3.8 provides the same results for functional damage to vehicle. The relationship with Pr(Injury) is ambiguous, which may reflect categorical coding of vehicle damage severity.

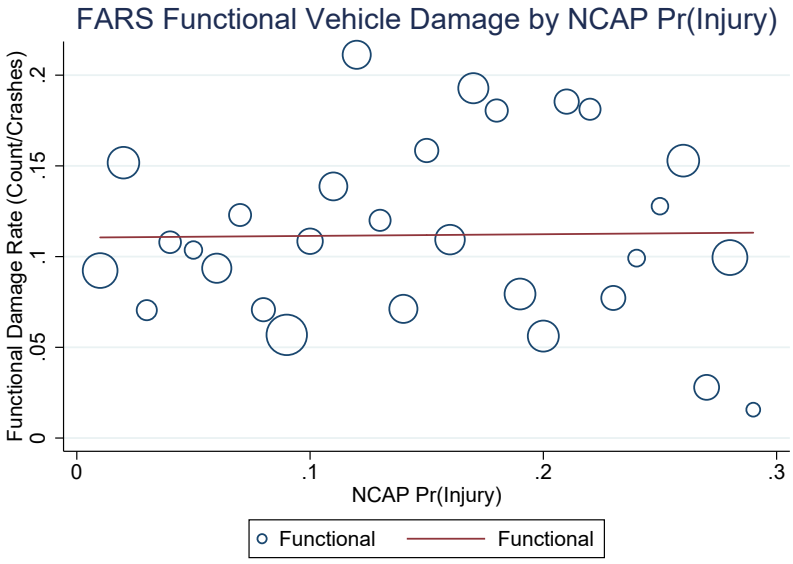


Figure 3.8: FARS Vehicle Damage: Functional

Figure 3.9 plots rate of minor vehicle damage against Pr(Injury). There is a negative correlation, as we would expect, as "minor" is the least severe category of vehicle damage, excluding no damage.

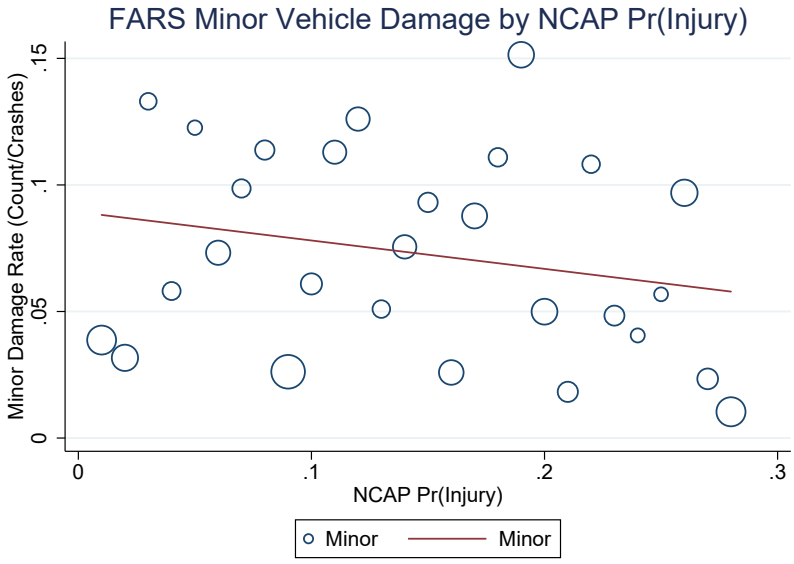


Figure 3.9: FARS Vehicle Damage: Minor

Because vehicle damage is coded categorically in the FARS data, no damage rate (the inverse of "any" damage) may reveal a stronger correlation with NCAP Pr(Injury). Figure 3.10 plots the rate of no vehicle damage against Pr(Injury). The relationship is ambiguous.

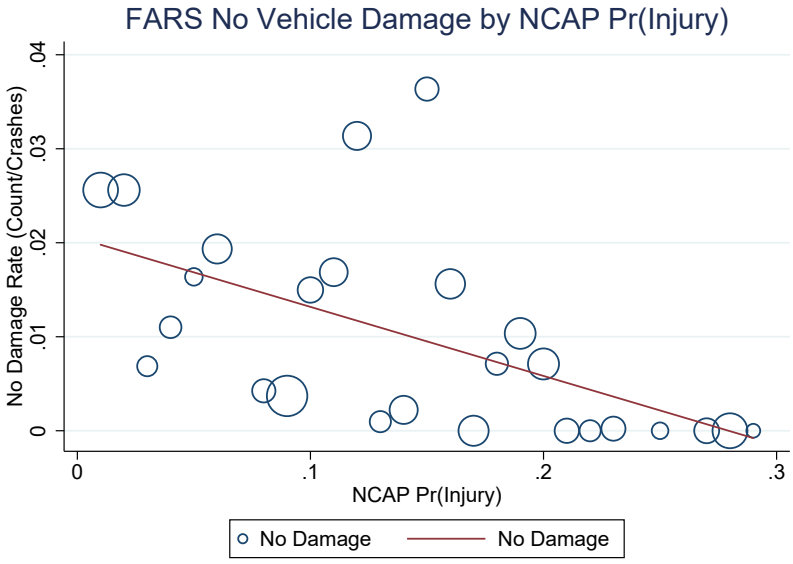


Figure 3.10: FARS Vehicle Damage: None

Figure 3.11 plots no injury rate against NCAP Pr(Injury). The relationship is negative, as we would expect.

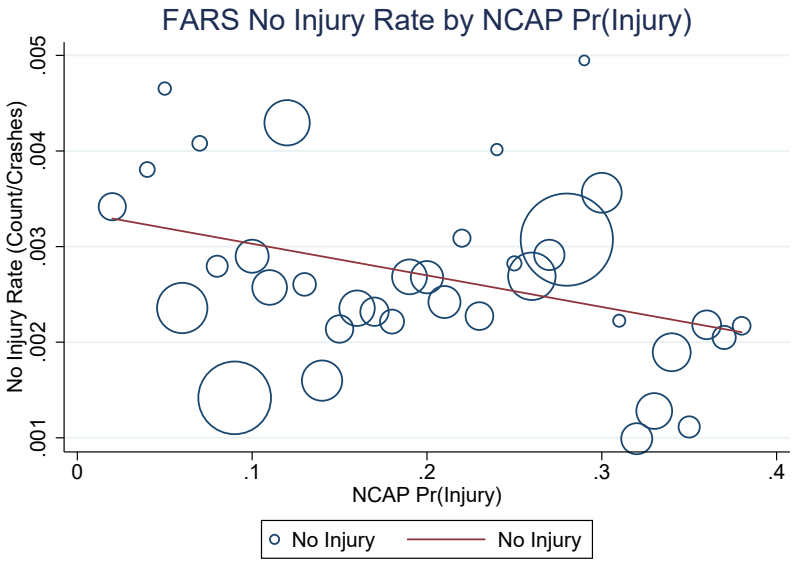


Figure 3.11: FARS No Injury

Figure 3.12 provides the same figure for possible injury. The relationship is weakly negative.

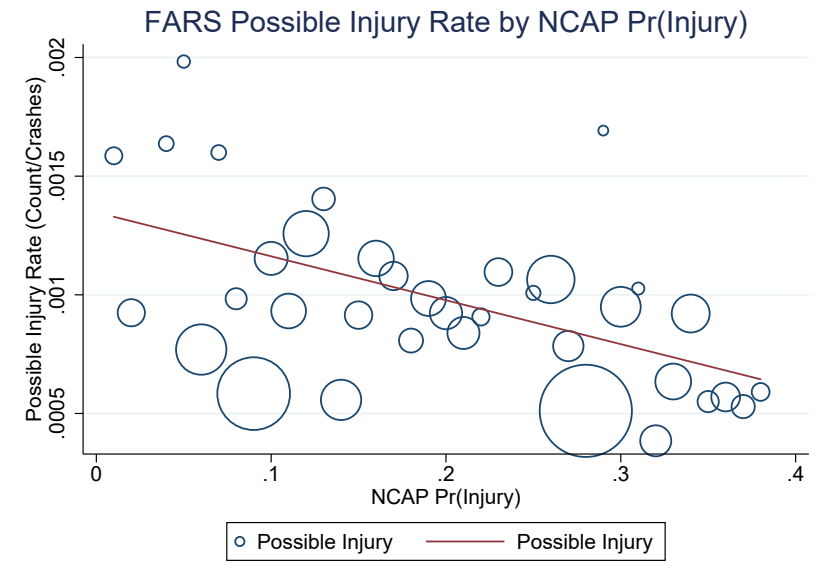


Figure 3.12: FARS Possible Injury

Figure 3.13 provides the results for non-incapacitating injury rate. The relationship is ambiguous.

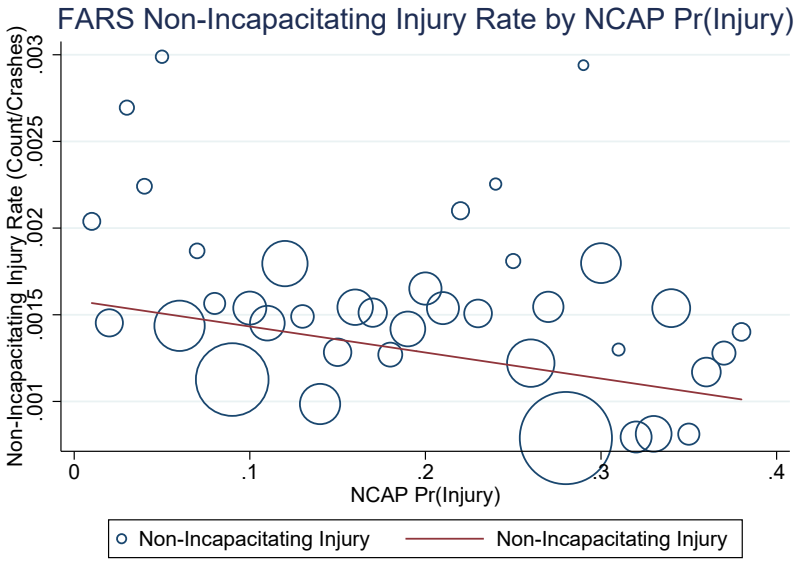


Figure 3.13: FARS Non-Incapacitating Injury

Figure 3.14 provides the results for incapacitating injury. I would expect a strong positive correlation with NCAP Pr(Injury); however, the relationship is ambiguous.

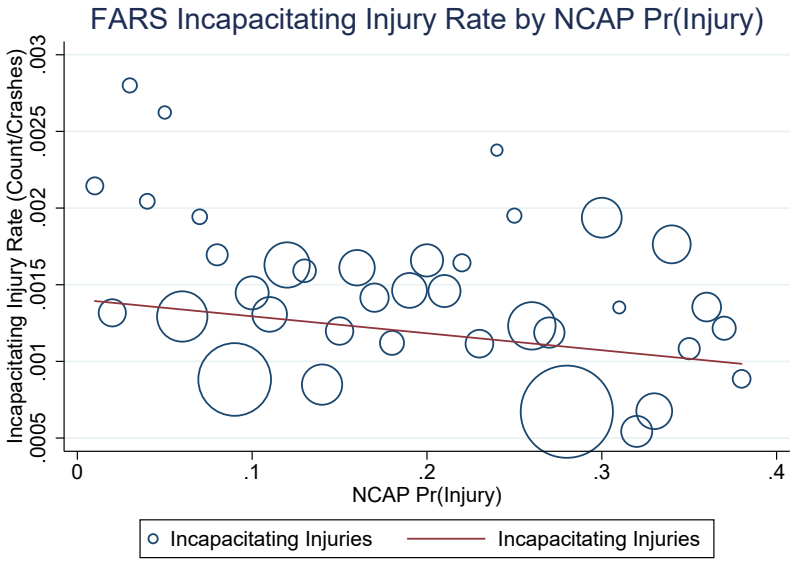


Figure 3.14: FARS Incapacitating Injury

Finally, figure 3.15 plots fatal injury rate against NCAP Pr(Injury). As predicted, there is a strong positive relationship.

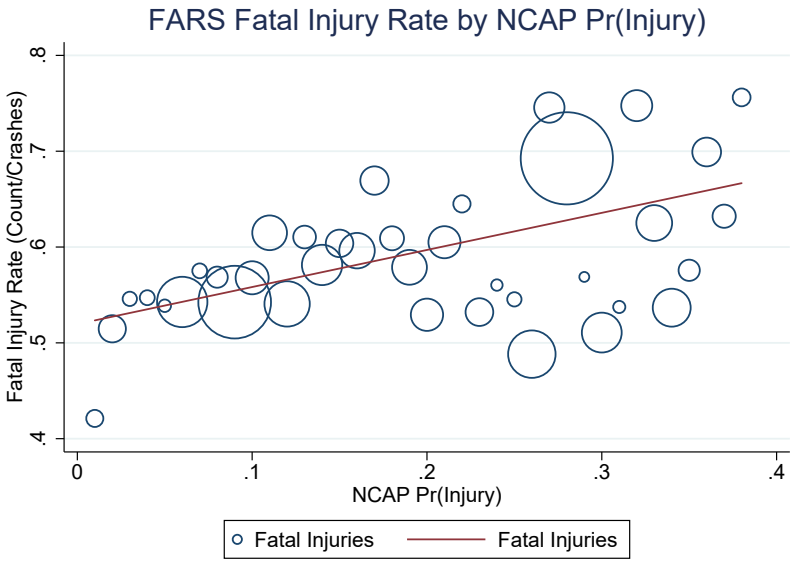


Figure 3.15: FARS Fatal Injury

Table 3.3 provides estimates for fatal injury outcomes. Reported estimates are based on MY1994-2010 vehicles for all star ratings and test types. I cluster at the vehicle model level and weight by total crash counts. Injury outcomes are weighted by number of vehicle occupants. My estimates are not sensitive to the inclusion of the quadratic, so my proceeding results will present linear regression estimates.

VARIABLES	(1) Fatal Injury	(2) Fatal Injury	(3) Fatal Injury	(4) Fatal Injury
NCAP Pr(Injury)	0.00196**	0.00317*	0.000505	-0.00116
	(0.000828)	(0.00171)	(0.00202)	(0.00173)
Pr(Injury)_sq		-0.00238	5.73e-05	0.00238
		(0.00218)	(0.00255)	(0.00253)
Ln(Curb Weight(lbs.))			-0.00153***	-0.00128**
			(0.000232)	(0.000595)
Ln(Max HP)				0.000113
				(0.000279)
Ln(MPG)(city)				-0.00173**
				(0.000719)
Ln(MPG)(highway)				0.00252***
				(0.000889)
Ln(Speed)			3.54e-05	3.66e-05
			(4.13e-05)	(5.50e-05)
Driver accident within past 3 years			2.15e-05	3.52e-05
			(2.35e-05)	(2.75e-05)
Driver DWI in past 3 years			-3.73e-05**	-6.11e-05*
			(1.49e-05)	(3.45e-05)
Driver speeding in past 3 years			-9.75e-05	-0.000187**
			(7.55e-05)	(9.29e-05)
Constant	0.000489**	0.000386	0.0134***	0.00801
	(0.000233)	(0.000281)	(0.00215)	(0.00540)
Observations	206,208	206,208	187,815	139,908
R-squared	0.041	0.043	0.220	0.195

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.3: FARS Fatal Injury

Table 3.4 provides results for occupant injury outcomes. Reported estimates are for all test types and seat positions. The variation in injury rates appears to be driven by fuel efficiency. This is likely due to the strong negative correlation between fuel effieciency and curb weight. This relationship holds for all NCAP test types.

VARIABLES	(1) No Injury	(2) Possible	(3) Non-Incapacitating	(4) Incapacitating	(5) Fatal
NCAP Pr(Injury)	-0.000372	-0.000116	-0.000178	-0.000130	-8.08e-06
	(0.000394)	(0.000139)	(0.000220)	(0.000213)	(0.000574)
Ln(Curb Weight(lbs.))	-6.98e-05	-0.000104	-0.000102	-0.000169	-0.00129**
	(0.000450)	(0.000163)	(0.000254)	(0.000241)	(0.000604)
Ln(Max HP)	-0.000141	-3.78e-05	-0.000114	-0.000106	9.19e-05
	(0.000179)	(6.61e-05)	(9.58e-05)	(9.61e-05)	(0.000283)
Ln(MPG)(city)	-0.000882**	-0.000336**	-0.000638***	-0.000695***	-0.00176**
	(0.000398)	(0.000149)	(0.000200)	(0.000204)	(0.000707)
Ln(MPG)(highway)	0.00151***	0.000517***	0.000922***	0.000931***	0.00250***
	(0.000515)	(0.000194)	(0.000295)	(0.000286)	(0.000876)
Ln(Speed)	8.56e-05	3.36e-05	5.22e-05	5.28e-05	4.26e-05
	(6.18e-05)	(2.34e-05)	(3.60e-05)	(3.61e-05)	(6.02e-05)
Driver accident within past 3 years	5.47e-05	1.96e-05	3.13e-05	3.27e-05*	3.93e-05
	(3.39e-05)	(1.28e-05)	(1.94e-05)	(1.98e-05)	(3.10e-05)
Driver DWI in past 3 years	-3.31e-05	-1.37e-05	-1.95e-05	-1.44e-05	-6.07e-05*
	(2.33e-05)	(8.48e-06)	(1.25e-05)	(1.09e-05)	(3.41e-05)
Driver speeding in past 3 years	-0.000190*	-7.21e-05*	-0.000111*	-0.000107*	-0.000195*
	(9.91e-05)	(3.74e-05)	(5.77e-05)	(5.64e-05)	(0.000101)
Constant	-0.000632	0.000471	0.000486	0.00110	0.00821
	(0.00401)	(0.00142)	(0.00225)	(0.00209)	(0.00552)
Observations	139,908	139,908	139,908	139,908	139,908
R-squared	0.107	0.066	0.098	0.119	0.194

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.4: FARS Injury Outcomes

Table 3.5 breaks out the linear regression results for fatal injury rates by test type. Columns (1)-(6) provide estimates for Frontal Driver, Frontal Passenger, Side Driver, Side Passenger, Rollover and Rollover 4WD test results, respectively. The frontal and side driver regressions restricts to occupants in the driver position, while the front and side passenger regressions restricts to occupants in the front passenger position. The rollover and rollover 4WD regressions include all occupants.

Table 3.6 provides fatal injury results by vehicle type. The coefficient on Pr(Injury) is insignificant for all vehicle types. The negative relationship between fatality rate and curb weight holds and is significant for Trucks and SUVs, the heavier vehicle types.

VARIABLES	(1) FD	(2) FP	(3) SD	(4) SP	(5) Roll	(6) Roll 4WD
NCAP Pr(Injury)	0.00114	0.000931*	-0.00270**	-6.89e-05	0.00286	0.0118***
	(0.000959)	(0.000520)	(0.00121)	(0.00125)	(0.00393)	(0.00307)
Ln(Curb Weight(lbs.))	-0.00214***	-0.000684	-0.00494***	-0.00146	-0.00288**	-0.00236**
	(0.000568)	(0.000428)	(0.00155)	(0.00104)	(0.00130)	(0.00106)
Ln(Max HP)	0.000383	2.69e-05	0.000985	0.000387	0.000838*	0.000524**
	(0.000406)	(0.000254)	(0.000859)	(0.000527)	(0.000449)	(0.000258)
Ln(MPG)(city)	-0.00292***	-0.000248	-0.00338	-0.00266	6.63e-05	0.00192
	(0.000954)	(0.000793)	(0.00250)	(0.00178)	(0.00140)	(0.00130)
Ln(MPG)(highway)	0.00291***	0.000951	0.00253	0.00154	0.00182	-0.00142
	(0.00108)	(0.000676)	(0.00260)	(0.00155)	(0.00112)	(0.000925)
Ln(Speed)	-4.41e-05*	-4.44e-05*	6.82e-06	1.78e-05	-3.05e-05	2.99e-05***
	(2.26e-05)	(2.46e-05)	(3.16e-05)	(3.36e-05)	(3.63e-05)	(8.35e-06)
Driver accident	2.45e-05	-4.62e-05*	-3.08e-05	8.42e-06	2.45e-05	6.45e-05**
within past 3 years	(3.88e-05)	(2.39e-05)	(6.64e-05)	(4.77e-05)	(2.36e-05)	(3.01e-05)
Driver DWI	-0.000138	0.000150**	0.000266*	8.71e-05	-5.15e-05***	-2.48e-05
within past 3 years	(0.000105)	(6.27e-05)	(0.000135)	(0.000136)	(1.86e-05)	(2.31e-05)
Driver speeding	-1.19e-05	1.86e-05	5.56e-05	-2.11e-05	-1.95e-05	-2.13e-06
within past 3 years	(3.68e-05)	(2.56e-05)	(5.30e-05)	(5.32e-05)	(3.11e-05)	(1.92e-05)
Constant	0.0160***	0.00400	0.0388**	0.0134	0.0144	0.0142*
	(0.00470)	(0.00312)	(0.0153)	(0.0108)	(0.00932)	(0.00769)
Observations	4,642	1,355	14,949	3,046	6,560	2,579
R-squared	0.228	0.219	0.122	0.023	0.260	0.089

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.5: FARS Fatalities by NCAP Crash Test Type

Table 3.7 shows my regression results for the FARS data for vehicle damage outcomes. Vehicle damage rates are calculated based on total crash count by model. There does not appear to be a strong relationship with NCAP Pr(Injury). Restricting results to frontal collisions and by vehicle type does not alter my findings.

VARIABLES	(1) Cars	(2) Trucks	(3) SUVs	(4) Vans
NCAP Pr(Injury)	-0.00143	0.000657	0.000259	-0.000884
	(0.00101)	(0.000980)	(0.00223)	(0.00324)
Ln(Curb Weight(lbs.))	0.00387	-0.00339**	-0.00438**	-0.00437
	(0.00254)	(0.00121)	(0.00207)	(0.00721)
Ln(Max HP)	0.00142	0.000801	0.00432***	-0.000861
	(0.00112)	(0.000511)	(0.00124)	(0.00195)
Ln(MPG)(city)	0.00746**	0.000809	0.0116***	-0.0199
	(0.00362)	(0.00105)	(0.00400)	(0.0124)
Ln(MPG)(highway)	-0.00442	-0.00108	-0.00408	0.00898
	(0.00304)	(0.00131)	(0.00361)	(0.00808)
Ln(Speed)	-2.22e-05	-3.30e-05*	-0.000404**	-2.96e-05
	(8.06e-05)	(1.89e-05)	(0.000162)	(9.84e-05)
Driver accident within past 3 years	3.37e-05	-3.28e-05	0.000192*	0.000113
	(6.53e-05)	(3.08e-05)	(9.62e-05)	(0.000234)
Driver DWI in past 3 years	2.25e-05	-4.07e-05	-2.92e-05	0.000191
	(0.000140)	(4.90e-05)	(0.000299)	(0.000803)
Driver speeding in past 3 years	-0.000248	-1.88e-05	-1.62e-05	0.000393
	(0.000161)	(1.43e-05)	(0.000123)	(0.000365)
Constant	-0.0430*	0.0259**	-0.00120	0.0715
	(0.0230)	(0.0101)	(0.0217)	(0.0657)
Observations	80,296	29,828	21,913	7,871
R-squared	0.016	0.102	0.035	0.006

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.6: FARS Outcomes by Vehicle Type

3.5.2 TX CRIS

Figure 3.16 plots the distribution of NCAP Pr(Injury) for the Texas CRIS data. I restrict my results for the CRIS data to $Pr(Injury) \leq 0.25$.

VARIABLES	(1) Crashes	(2) Deaths	(3) Minor	(4) Functional	(5) Disabling	(6) No Damage
NCAP Pr(Injury)	65.55	0.0343	0.326	-0.294	-0.130	0.0260
	(67.52)	(0.0654)	(0.243)	(0.285)	(0.354)	(0.0260)
Ln(Curb Weight(lbs.))	12.79	-0.602***	0.112	-0.607*	0.636*	0.00320
	(73.62)	(0.109)	(0.106)	(0.326)	(0.358)	(0.0117)
Ln(Max HP)	-81.07*	0.119**	0.00300	0.375**	-0.375**	-0.00195
	(46.63)	(0.0478)	(0.0551)	(0.153)	(0.166)	(0.00773)
Ln(MPG)(city)	-154.6*	-0.271	0.434**	0.527*	-0.682**	-0.0648
	(92.07)	(0.165)	(0.177)	(0.276)	(0.335)	(0.0467)
Ln(MPG)(highway)	180.5*	0.270**	-0.387**	-0.778**	0.944**	0.0646
	(92.61)	(0.120)	(0.183)	(0.338)	(0.422)	(0.0469)
Ln(Speed)	6.061***	-0.00471	0.0147**	-0.0367	0.0157	-0.000162
	(2.301)	(0.0105)	(0.00668)	(0.0297)	(0.0238)	(0.000789)
Driver accident	-1.265	-0.0126**	-0.0105	-0.00799	0.0193	-0.00162
within past 3 years	(3.222)	(0.00577)	(0.0112)	(0.0170)	(0.0190)	(0.00104)
Driver DWI	13.86*	-0.00104	0.0392**	-0.0215	-0.0266	-0.00249
within past 3 years	(7.811)	(0.00687)	(0.0193)	(0.0193)	(0.0251)	(0.00162)
Driver speeding	7.766***	0.0196	0.000744	0.0386	-0.0367	-0.00261*
within past 3 years	(2.433)	(0.0170)	(0.00775)	(0.0503)	(0.0393)	(0.00153)
Constant	246.4	4.837***	-0.981	4.343	-3.633	-0.0333
	(577.6)	(0.997)	(1.078)	(2.974)	(3.271)	(0.141)
Observations	67,277	142,133	142,133	142,133	142,133	142,133
R-squared	0.142	0.543	0.031	0.041	0.028	0.005

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.7: FARS Vehicle Damage

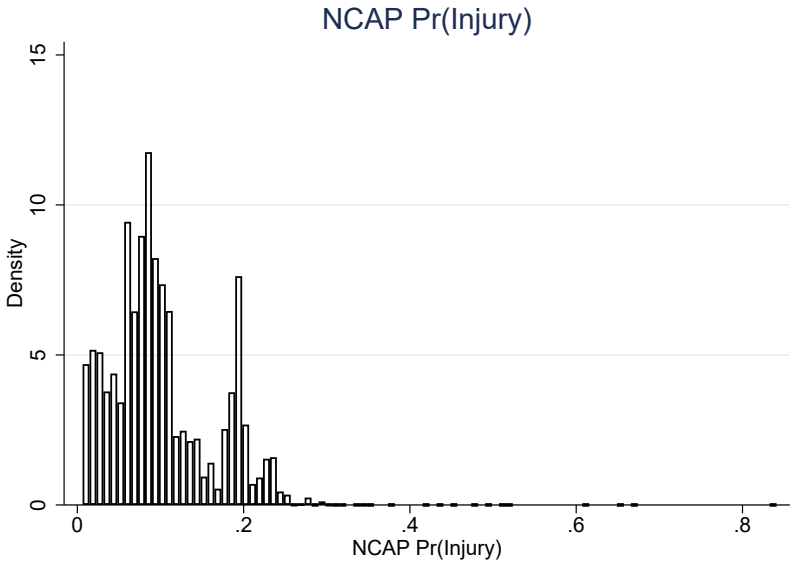


Figure 3.16: NCAP Pr(Injury) Distribution: TX

Figure 3.17 shows CRIS fatality rate by NCAP Pr(Injury). Rates are total counts over total crashes by model. There is a positive correlation between NCAP Pr(Injury) and fatality rate, as we would expect.

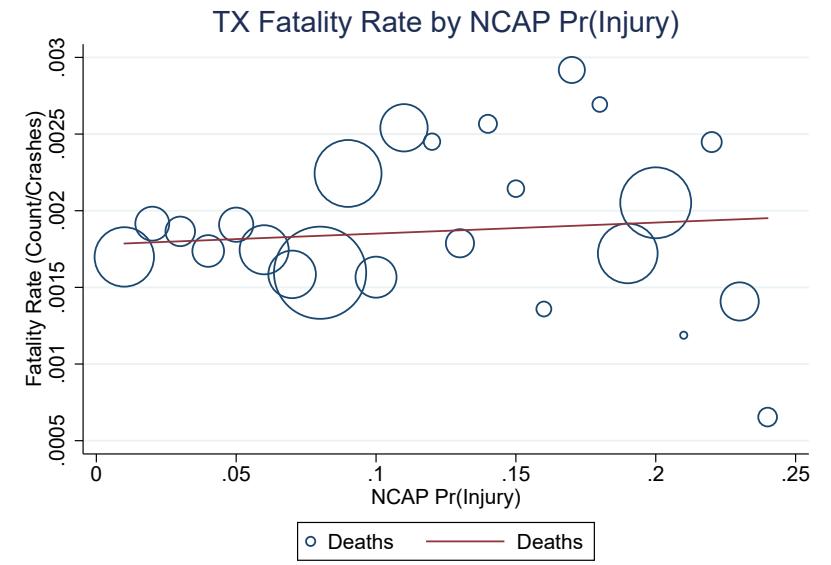


Figure 3.17: TX Fatalities

Figure 3.18 shows the same figure for incapacitating injury rate.The relationship is positive, as expected.

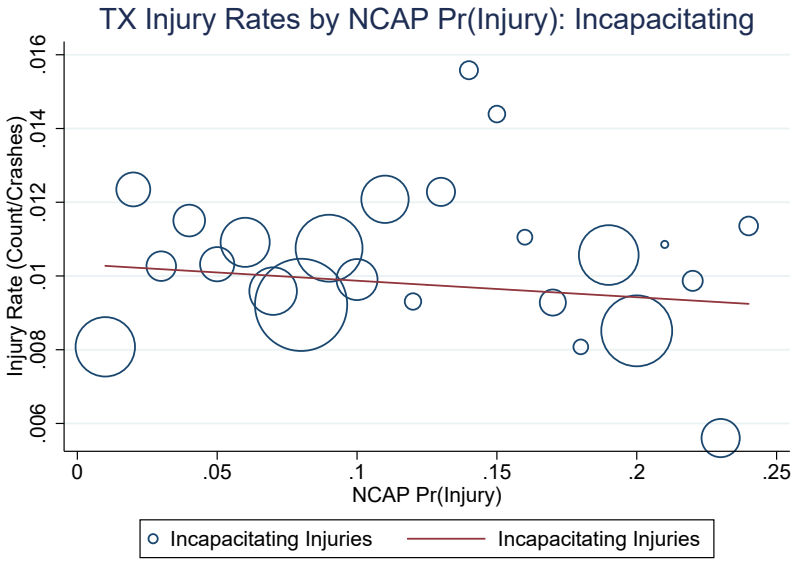


Figure 3.18: TX Injuries: Incapacitating

Figure 3.19 shows the same results for non-incapacitating injury. The positive relationship between NCAP Pr(Injury) and crash vehicle occupant injury rate holds.

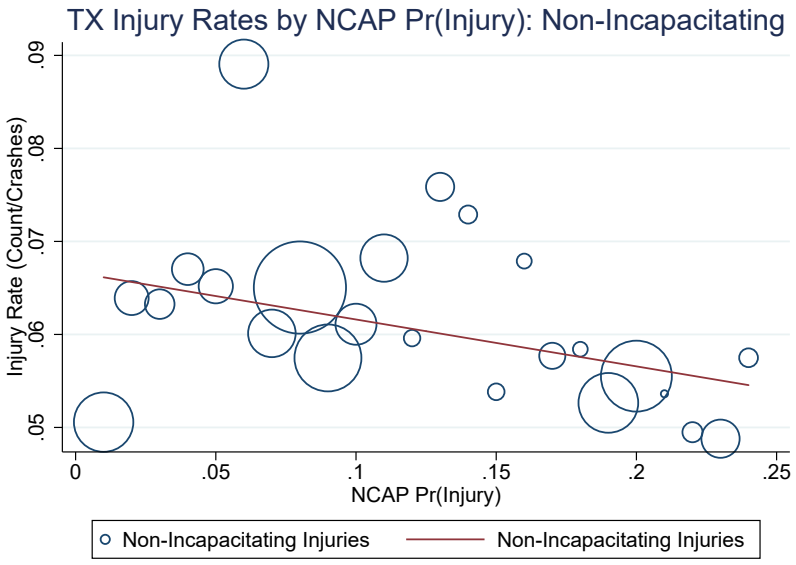


Figure 3.19: TX Injuries: Non-Incapacitating

Figure 3.20 shows the results for "possible" injury. There is a negative correlation with NCAP Pr(Injury). This is likely due to categorical coding occupant injury severity.

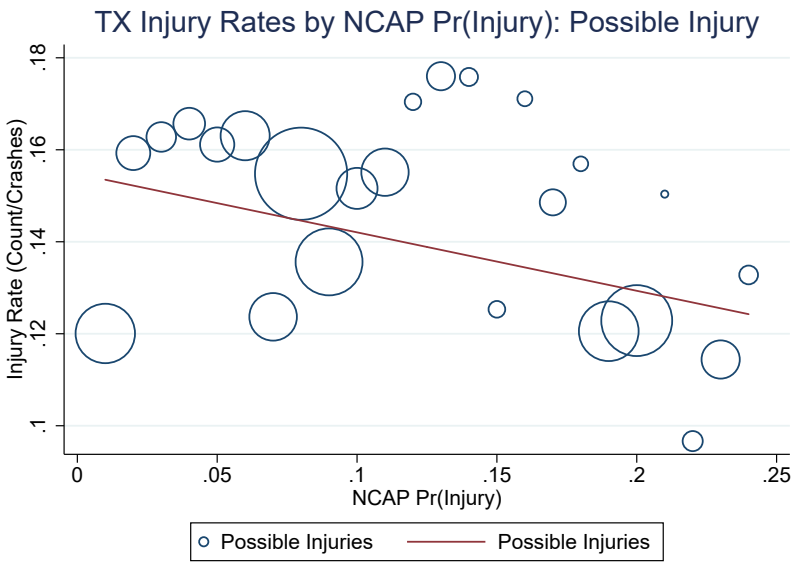


Figure 3.20: TX Injuries: Possible

In order to overcome categorical coding of occupant injury severity, I consider mean total number of occupant injuries per vehicle in the CRIS data. Figure 3.21 shows total injury rate by vehicle model plotted against NCAP Pr(Injury). As expected, there is a positive correlation: less safe vehicles experience higher rates of occupant injury.

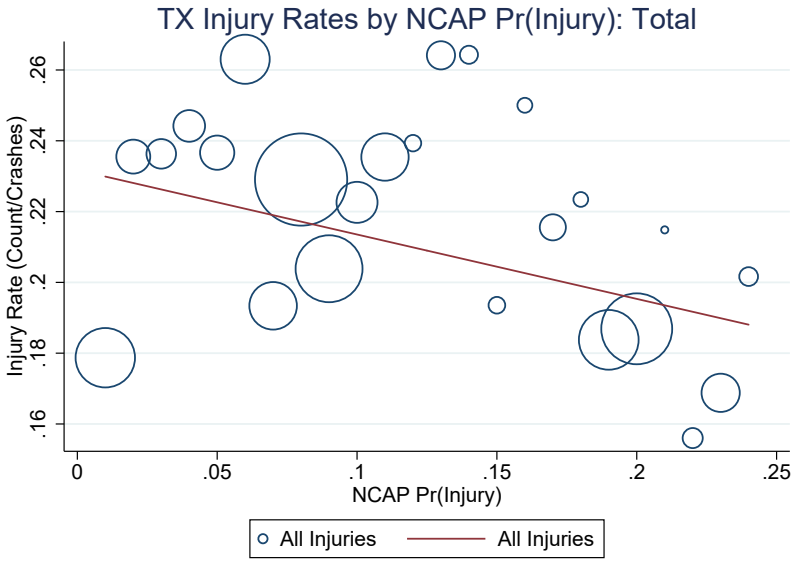


Figure 3.21: TX Injuries: Total

Figure 3.22 shows the same figure for no injury rate. As predicted, there is a negative correlation.

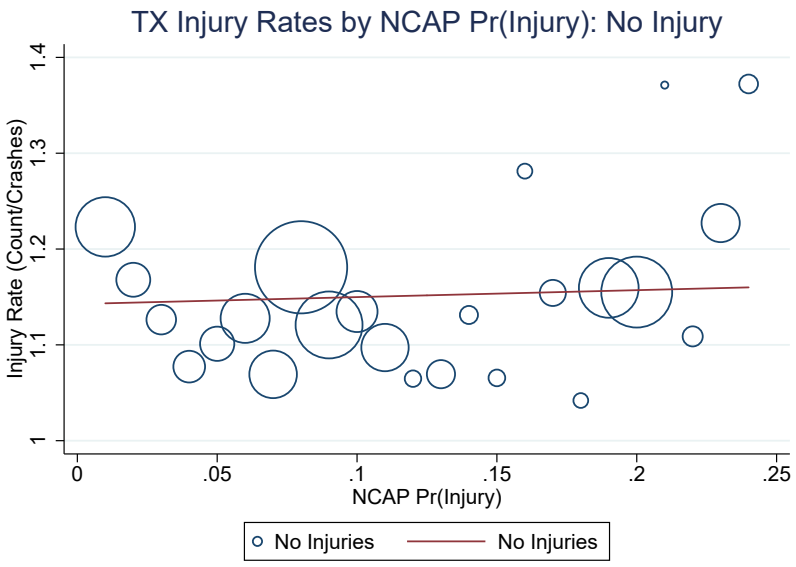


Figure 3.22: TX Injuries: None

Figure 3.23 plots vehicle damage against NCAP Pr(Injury). Vehicle damage is rated on a scale of 1 = minimal damage to 7 = maximum damage. The relationship is ambiguous.

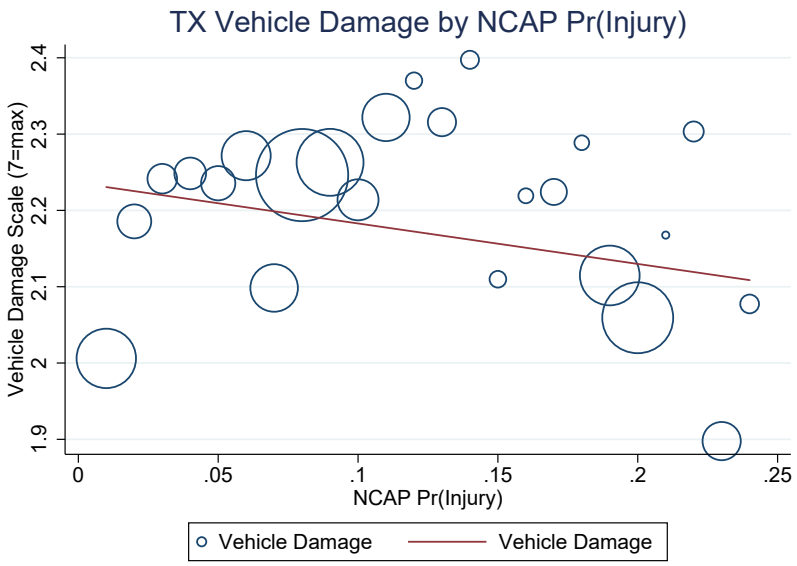


Figure 3.23: TX Vehicle Damage

Figure 3.24 shows that there is no evidence of discontinuity in total crash count in the TX CRIS data at the NCAP star ratings threshold. There is therefore not evidence strictly from crash count numbers of driver selection, e.g. safer drivers purchasing "safer" (e.g. 5- vs. 4-star) vehicles and driving them more safely.

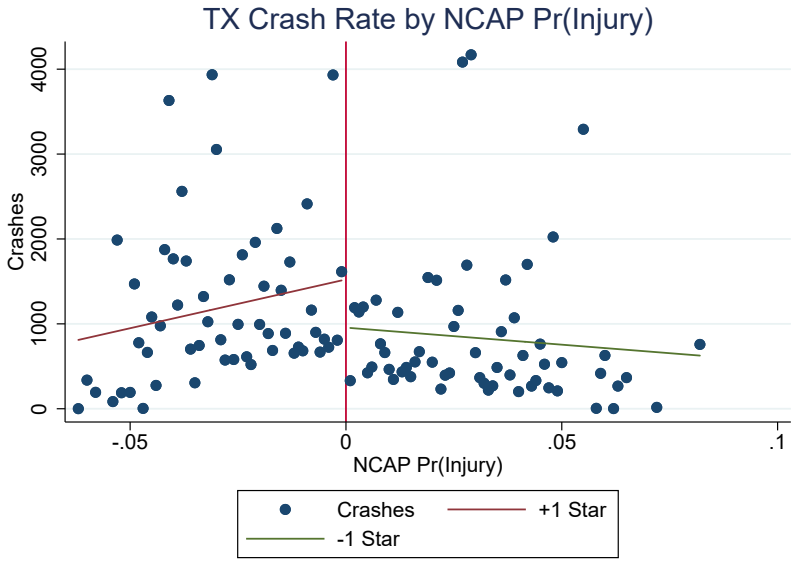


Figure 3.24: TX Crashes vs. NCAP Pr(Injury) (pooled)

Table 3.8 shows my RD estimates for the Texas CIRS data for injury and fatality outcomes. Outcomes are total crash counts by model, and fatality and injury rates as total counts over total crashes by model. The large negative coefficient for NCAP Pr(Injury) on total crashes is due to the fact that this is total crash count and does not reflect sales volumes, which are decreasing in NCAP Pr(Injury). However, the relationship between Pr(Injury) and occupant injury outcomes is not significant. This is likely due to the high correlation between Pr(Injury) and curb weight. There is a significant negative correlation between vehicle curb weight and fatality and total injury rates.

Table 3.9 provide my results for the Texas RIS data for vehicle damage outcomes. Rates are counts over total crash counts within 2 years of modey year by model.

VARIABLES	(1) Total Crashes	(2) Deaths	(3) Total Injuries	(4) Incapacitating	(5) Non-Incapacitating
NCAP Pr(Injury)	-1,387	0.00192	-0.00600	0.000111	-0.00380
	(1,667)	(0.00156)	(0.0184)	(0.00416)	(0.00786)
Ln(Curb Weight)	-66.56	-0.00270**	-0.104***	-0.0131***	-0.0476***
	(1,606)	(0.00130)	(0.0244)	(0.00325)	(0.0130)
Ln(MPG)(city)	-2,677	-0.00202	-0.0595*	-0.00394	-0.0139
	(1,870)	(0.00269)	(0.0357)	(0.00297)	(0.0117)
Ln(MPG)(highway)	4,299	-0.00277	0.102***	-0.00259	0.00252
	(2,810)	(0.00209)	(0.0363)	(0.00371)	(0.0138)
Ln(Max HP)	551.6	-0.000807	-0.0223	0.000326	-0.00312
	(569.9)	(0.00170)	(0.0202)	(0.00192)	(0.00932)
Constant	-6,672	0.0432**	1.042***	0.137***	0.504***
	(14,489)	(0.0204)	(0.189)	(0.0318)	(0.0987)
Observations	120,244	164,148	164,148	164,148	164,148
R-squared	0.205	0.082	0.637	0.187	0.161

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.8: TX Occupant Injury Outcomes

Damage is rated on a scale from 0 (no damage) to 7 (maximum damage). The relationship between vehicle damage and NCAP Pr(Injury) is ambiguous, but there is a strong negative relationship between vehicle damage and curb weight, reflecting the greater resilience of heavier vehicles in collisions.

Table 3.10 breaks out the linear regression results for fatal injury rates by test type. Columns (1)-(6) provide estimates for Frontal Driver, Frontal Passenger, Side Driver, Side Passenger, Rollover and Rollover 4WD test results, respectively. The correlation with NCAP Pr(Injury) is negative and significant as expected for the Rollover tests, but goes in the opposite direction to that we would expect for the Frontal Passenger test. The results for the other crash test types are ambiguous.

Table 3.11 provides the same results for total injury rates. The correlation between NCAP Pr(Injury) and total occupant injury rate is weak and ambiguous. However, there is a strong negative relationship between total occupant injury and curb

VARIABLES	(1) Damage	(2) Damage	(3) Damage	(4) Damage
NCAP Pr(Injury)	-0.610***	2.389***	-0.00170	0.258
	(0.153)	(0.808)	(0.182)	(0.712)
Ln(Curb Weight)			-0.670***	-0.664***
			(0.138)	(0.136)
Ln(MPG)(city)			0.0997	0.105
			(0.191)	(0.198)
Ln(MPG)(highway)			-0.119	-0.123
			(0.136)	(0.137)
Ln(Max HP)			0.0258	0.0258
			(0.135)	(0.134)
Pr(Injury)*2		-12.32***		-1.106
		(3.684)		(2.562)
Constant	2.195***	2.067***	7.695***	7.628***
	(0.0563)	(0.0388)	(1.720)	(1.777)
Observations	172,277	172,277	160,773	160,773
R-squared	0.000	0.001	0.005	0.005

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.9: TX Vehicle Damage Outcomes (max=7)

VARIABLES	(1) FD	(2) FP	(3) SD	(4) SP	(5) Roll	(6) Roll 4WD
NCAP Pr(Injury)	0.0141	0.0157**	-0.00414	-0.00738	-0.0183**	-0.0329**
	(0.00965)	(0.00663)	(0.00868)	(0.00615)	(0.00720)	(0.0150)
Ln(Curb Weight)	-0.00226	-0.000368	-0.00340*	-0.00236	-0.00394*	0.00347
	(0.00232)	(0.000855)	(0.00176)	(0.00169)	(0.00216)	(0.00262)
Ln(MPG)(city)	-0.00443	-0.00147	-0.000996	-0.00167	-0.00496	-0.00456
	(0.00557)	(0.00242)	(0.00220)	(0.00118)	(0.00351)	(0.00568)
Ln(MPG)(highway)	-0.00169	-0.00143	-0.000931	-0.000533	-0.00457	-0.0118**
	(0.00295)	(0.00205)	(0.00316)	(0.00220)	(0.00287)	(0.00470)
Ln(Max HP)	-0.00229	-0.000506	0.000520	-0.000244	-0.00108	-0.00844***
	(0.00221)	(0.00100)	(0.00150)	(0.000944)	(0.00157)	(0.00301)
Constant	0.0503*	0.0148	0.0331**	0.0295*	0.0721***	0.0750**
	(0.0295)	(0.0100)	(0.0137)	(0.0150)	(0.0248)	(0.0308)
Observations	35,364	38,313	23,759	20,221	34,556	11,935
R-squared	0.142	0.163	0.035	0.042	0.228	0.460

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.10: TX Fatalities by NCAP Test Type

weight, as expected.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	FD	FP	SD	SP	Roll	Roll 4WD
NCAP Pr(Injury)	-0.112	0.0876	-0.149	-0.0176	0.00608	-0.209***
	(0.193)	(0.108)	(0.133)	(0.0690)	(0.0845)	(0.0746)
Ln(Curb Weight)	-0.184**	-0.0702*	-0.123***	-0.166***	-0.0527*	-0.0377*
	(0.0752)	(0.0410)	(0.0306)	(0.0266)	(0.0310)	(0.0227)
Ln(MPG)(city)	-0.0487	-0.0719	-0.0786**	-0.0475*	-0.0731	0.0899***
	(0.0349)	(0.0864)	(0.0316)	(0.0283)	(0.0605)	(0.0301)
Ln(MPG)(highway)	0.0720	0.105	0.0779	0.0474	0.149**	0.0138
	(0.0506)	(0.0819)	(0.0528)	(0.0456)	(0.0737)	(0.0283)
Ln(Max HP)	0.0177	-0.0440	-0.0369**	-0.0254	-0.0477**	0.0188**
	(0.0470)	(0.0336)	(0.0176)	(0.0206)	(0.0192)	(0.00894)
Constant	1.559***	0.888***	1.413***	1.714***	0.639**	0.142
	(0.535)	(0.250)	(0.314)	(0.310)	(0.321)	(0.175)
Observations	35,364	38,313	23,759	20,221	34,556	11,935
R-squared	0.428	0.699	0.625	0.814	0.787	0.589

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.11: TX Occupant Injuries by NCAP Test Type

3.6 Conclusion

Publication of safety information can provide a positive public externality, but any positive effects will be distorted if there are informational inaccuracies. This paper evaluates whether NHTSA's New Car Assessment Program vehicle safety ratings accurately reflect real-world injury and fatality rates. I match NCAP safety ratings with real-world crash data from the Fatal Accident Reporting System (FARS) and Texas' Crash Records Information System (CRIS). I find minimal correlation between real-world crash outcomes and predicted risk of injury calculated in simulated NCAP crash tests. My findings suggest reflect significant differences between labratory crash testing procedures and actual driving conditions. My results call into question the ultimate value of the NCAP information program to consumers.

In this paper, I am unable to address concerns of compensating behavior by drivers which may distort intended policy effects of the NCAP program. For example, drivers of "safer"-rated vehicles may drive more recklessly, i.e. moral hazard. [62] Future work will use vehicle registration data to control for vehicle miles driven and additional driver characteristics.

In addition, in future work I would also like to address how design standards and regulations around consumer preferences and product market competition can be used as a public policy tool to promote public health and safety on the roads. The auto industry has historically been a driving force in the U.S. economy, reflecting a strong manufacturing sector and consumer confidence. Further, the automobile manufacturing

industry is a primary laboratory for technological innovation, with advancements in vehicle technology having applications in medicine, education, communication, etc. At the same time, we observe vehicle and brand differentiation, signaling and other forms of monopolistic competition that suggest a potential role of policy intervention to promote innovation in vehicle safety technology, and the adoption and the diffusion of these potentially life-saving technologies.

One goal of the NCAP program is to promote innovation in vehicle safety technology NCAP does not mandate safety technology, but producers may respond to a low rating by re-designing vehicles to perform better on a consecutive crash test. As a result, to the extent that NCAP crash test environment does not accurately reflect real-world driving conditions, this may lead to a distortionary effect in terms of manufacturer incentive to invest in technology that leads to better test performance at the expense of real-world safety. If NCAP affects producer choice over investment in safety technology, then any failures in the NCAP ratings to predict real-world vehicle performance will be borne out in injury and loss of life on a national scale.

These concerns, coupled with my results from Chapter 2 and this paper, raise is the question of NCAP reform. Should evaluation of safety ratings relative to real-world crashes be restricted to specific crash types similar to those of the testing conditions? Or do the ratings formulae need to be revised to better reflect real-world driving conditions? Any reform to NCAP will have widespread public health consequences, as the US NCAP program is emulated internationally. In addition to the US NCAP, there exist Australasian NCAP, Euro NCAP, C-NCAP (China), JNCAP (Japan), Korea-NCAP,

and others. Further, NHTSA's NCAP safety ratings program is only one form of federal vehicle regulation in the US. Perhaps other regulatory pathways would be more effective in promoting vehicle safety with lower risk of distorting manufacturer incentives and driver behavior.

Bibliography

[1] Achievement gaps: How hispanic and white students in public schools perform in mathematics and reading on the national assessment of educational progress. *U.S. Department of Education, NCES 2011.*

[2] Technical report for the california english language development test (celdt), 2000-2001. *Technical Report*, 2001.

[3] Frequently asked questions (faqs) for els in california. *California Department of Education*, 2006.

[4] National clearinghouse for english language acquisition and language instruction educational programs. state focus: California, school year 2009-2010. *National Clearinghouse for English Language Acquisition*, 2010.

[5] Title iii biennial report to congress, school years 2008-2010. state profiles: California. *U.S. Department of Education, Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students*, 2013.

[6] National Highway Traffic Safety Administration. Federal register. *Consumer Information; New Car Assessment Program*, 2008.

[7] Philippe Aghion, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728, 2005.

[8] Geokge A Akerlof. The market for" lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[9] Iliana Alanis. A texas two-way bilingual program: Its effects on linguistic and academic achievement. *Bilingual Research Journal*, 24(3):225–248, 2000.

[10] Michael Anderson and Jeremy Magruder. Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989, 2012.

[11] Joshua D Angrist and Alan B Krueger. Empirical strategies in labor economics. *Handbook of labor economics*, 3:1277–1366, 1999.

[12] Scott E. Atkinson and Robert Halvorsen. The valuation of risks to life: Evidence from the market for automobiles. *The Review of Economics and Statistics*, 72(1):133–136, 1990.

[13] W Steven Barnett, Donald J Yarosz, Jessica Thomas, Kwanghee Jung, and Dulce Blanco. Two-way and monolingual english immersion in preschool education: An

experimental comparison. *Early Childhood Research Quarterly*, 22(3):277–293, 2007.

[14] Hilary Bates, Matthias Holweg, Michael Lewis, and Nick Oliver. Motor vehicle recalls: Trends, patterns and emerging issues. *Omega*, 35(2):202–210, 2007.

[15] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[16] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[17] Mary A Burke and Tim R Sass. Classroom peer effects and student achievement. *Journal of Labor Economics*, 31(1):51–82, 2013.

[18] Rebecca M Callahan. Tracking and high school english learners: Limiting opportunity to learn. *American Educational Research Journal*, 42(2):305–328, 2005.

[19] BJ Campbell. A comparison of nhtsa car safety ratings with injuries in highway crashes. *University of North Carolina Highway Safety Research Center, Highway Safety Highlights*, 15(3), 1982.

[20] David Card, Carlos Dobkin, and Nicole Maestas. The impact of nearly universal insurance coverage on health care utilization and health: evidence from medicare. *American Economic Review*, 98(5):2242–2258, 2008.

[21] Scott E Carrell, Bruce I Sacerdote, and James E West. From natural variation to optimal policy? the lucas critique meets peer effects. Technical report, National Bureau of Economic Research, 2011.

[22] Injury Center. Data and statistics wisqars. *Overview*, Key Data and Statistics(Ten Leading Causes of Death and Injury):https://www.cdc.gov/injury/wisqars/LeadingCauses.html, 2017.

[23] Aimee Chin, N Meltem Daysal, and Scott A Imberman. Impact of bilingual education programs on limited english proficient students and their peers: Regression discontinuity evidence from texas. Technical report, National Bureau of Economic Research, 2012.

[24] Rosa Minhyo Cho. Are there peer effects associated with having english language learner (ell) classmates? evidence from the early childhood longitudinal study kindergarten cohort (ecls-k). *Economics of Education Review*, 31(5):629–643, 2012.

[25] Yoonhyeung Choi and Ying-Hsuan Lin. Consumer response to crisis: Exploring the concept of involvement in mattel product recalls. *Public Relations Review*, 35(1):18–22, 2009.

[26] Brian Cobb, Diego Vega, and Cindy Kronauge. Effects of an elementary dual language immersion school program on junior high school achievement. *Middle Grades Research: Exemplary Studies Linking Theory to Practice*, page 1, 2009.

[27] Michael Darden and Ian M McCarthy. The star treatment estimating the impact

of star ratings on medicare advantage enrollments. *Journal of Human Resources*, 50(4):980–1008, 2015.

[28] San Jose Unified School District. Parents: Enrollment process: Elementary – immediate attendance registration. 2013.

[29] Avinash K Dixit and Joseph E Stiglitz. Monopolistic competition and optimum product diversity. *The American economic review*, 67(3):297–308, 1977.

[30] Mark K Dreyfus and W Kip Viscusi. Rates of time preference and consumer valuations of automobile safety and fuel efficiency. *The Journal of Law and Economics*, 38(1):79–105, 1995.

[31] Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. Technical report, National Bureau of Economic Research, 2008.

[32] ED.gov. Ed data express: Data about elementary & secondary schools in the u.s. Technical report, U.S. Department of Education, 2013.

[33] Peggy Estrada. English learner curricular streams in four middle schools: Triage in the trenches. *The Urban Review*, 46(4):535–573, 2014.

[34] Charles M. Farmer. Relationships of frontal offset crash test results to real-world driver fatality rates. *Traffic Injury Prevention*, 6(1):31–37, 2005. PMID: 15823872.

[35] Patricia Gandara, Russell Rumberger, Julie Maxwell-Jolly, and Rebecca Callahan.

English learners in california schools: Unequal resources, unequal outcomes. *education policy analysis archives*, 11(36):1–54, 2003.

[36] Fred Genesee, Kathryn Lindholm-Leary, William Saunders, and Donna Christian. English language learners in us schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, 10(4):363–385, 2005.

[37] Fred Genesee, Kathryn Lindholm-Leary, William Saunders, and Donna Christian. English language learners in us schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, 10(4):363–385, 2005.

[38] Deborah C Girasek and Brett Taylor. An exploratory study of the relationship between socioeconomic status and motor vehicle safety features. *Traffic injury prevention*, 11(2):151–155, 2010.

[39] Gene M Grossman and Elhanan Helpman. Endogenous innovation in the theory of growth. *Journal of Economic Perspectives*, 8(1):23–44, 1994.

[40] Tamara Halle, Elizabeth Hair, Laura Wandner, Michelle McNamara, and Nina Chien. Predictors and outcomes of early versus later english language proficiency among english language learners. *Early childhood research quarterly*, 27(1):1–20, 2012.

[41] Laurie A Hellinga, Anne T McCartt, and Emily R Haire. Choice of teenagers' vehicles and views on vehicle safety: Survey of parents of novice teenage drivers. *Journal of safety research*, 38(6):707–713, 2007.

[42] Lawrence L Hershman. The us new car assessment program (ncap): Past, present and future. Technical report, SAE Technical Paper, 2001.

[43] Koichiro Ito and James M Sallee. The economics of attribute-based regulation: Theory and evidence from fuel economy standards. *Review of Economics and Statistics*, 100(2):319–336, 2018.

[44] Ginger Zhe Jin and Phillip Leslie. The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, 118(2):409–451, 2003.

[45] Ian S Jones and RA Whitfield. Predicting injury risk with new car assessment program crashworthiness ratings. *Accident Analysis & Prevention*, 20(6):411–419, 1988.

[46] Michael W Jones-Lee. The value of human life in the demand for safety: comment. *The American Economic Review*, 68(4):712–716, 1978.

[47] Charles J Kahane. Lives saved by vehicle safety technologies and associated federal motor vehicle safety standards, 1960 to 2012–passenger cars and ltvs–with reviews of 26 fmvss and the effectiveness of their associated safety technologies in reducing fatalities, injuries, and crashes. *Report No. DOT HS*, 812:069, 2015.

[48] Christopher R Knittel. Automobiles on steroids: Product attribute trade-offs and technological progress in the automobile sector. *American Economic Review*, 101(7):3368–99, 2011.

[49] Sjaanie Koppel, Judith Charlton, Brian Fildes, and Michael Fitzharris. How important is vehicle safety in the new vehicle purchase process? *Accident analysis & prevention*, 40(3):994–1004, 2008.

[50] M Krafft, A Kullgren, M Les, C Tingvall, and A Lie. Injury as a function of change of velocity-an alternative method to derive risk functions. In *VEHICLE SAFETY 2000-PROCEEDINGS OF AN INTERNATIONAL CONFERENCE HELD 7-9 JUNE 2000, INSTITUTION OF MECHANICAL ENGINEERS, LONDON, UK*, 2000.

[51] Anders Kullgren, Anders Lie, and Claes Tingvall. Comparison between euro ncap test results and real-world crash data. *Traffic Injury Prevention*, 11(6):587–593, 2010. PMID: 21128188.

[52] Teik Hua Law, Robert B Noland, and Andrew W Evans. The sources of the kuznets relationship between road fatalities and economic growth. *Journal of Transport Geography*, 19(2):355–365, 2011.

[53] Ethan G Lewis. Immigrant-native substitutability: the role of language ability. Technical report, National Bureau of Economic Research, 2011.

[54] Anders Lie and Claes Tingvall. How do euro ncap results correlate with real-life injury risks? a paired comparison study of car-to-car crashes. *Traffic Injury Prevention*, 3(4):288–293, 2002.

[55] Jordan D Matsudaira. Sinking or swimming? *Evaluating the impact of English immersion*, 2009.

[56] Peter McHenry and Melissa McInerney. Updated estimates of hispanic-white wage gaps for men and women. In *Proceedings from the American Economic Association Conference. Philadelphia, PA*, 2013.

[57] Kristina B. Metzger, Siobhan Gruschow, Dennis R. Durbin, and Allison E. Curry. Association between ncap ratings and real-world rear seat occupant risk of injury. *Traffic Injury Prevention*, 16(sup2):S146–S152, 2015. PMID: 26436224.

[58] Alfonso Miranda and Yu Zhu. English deficiency and the native–immigrant wage gap. *Economics Letters*, 118(1):38–41, 2013.

[59] D Mitchell, Tom Destino, and R Karan. Evaluation of english language development programs in the santa ana unified school district. *Riverside, CA: California Educational Research Cooperative, University of California, Riverside*, 1997.

[60] Stuart V Newstead, Sanjeev Narayan, Maxwell H Cameron, and Charles M Farmer. Us consumer crash test results and injury risk in police-reported crashes. *Traffic injury prevention*, 4(2):113–127, 2003.

[61] Thomas B Parrish, Amy Merickel, María Pérez, Robert Linquanti, Miguel Socias, Angeline Spain, Cecilia Speroni, Phil Esra, Leslie Brock, and Danielle Delancey. Effects of the implementation of proposition 227 on the education of english learn-

ers, k-12: Findings from a five-year evaluation. final report for ab 56 and ab 1116. *American Institutes For Research*, 2006.

[62] Sam Peltzman. The effects of automobile safety regulation. *Journal of political Economy*, 83(4):677–725, 1975.

[63] Nolan G Pope. The marginal effect of k-12 english language development programs: Evidence from los angeles schools. *Economics of Education Review*, 53:311–328, 2016.

[64] Sean F Reardon and Claudia Galindo. The hispanic-white achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, 46(3):853–891, 2009.

[65] Sean F Reardon and Joseph P Robinson. Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, 5(1):83–104, 2012.

[66] David A Reinstein and Christopher M Snyder. The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The journal of industrial economics*, 53(1):27–51, 2005.

[67] Joseph P Robinson. Evaluating criteria for english learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 33(3):267–292, 2011.

[68] Joseph P Robinson. Evaluating criteria for english learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 33(3):267–292, 2011.

[69] Paul M Romer. Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102, 1990.

[70] Nicholas G Rupp. Are government initiated recalls more damaging for shareholders? evidence from automotive recalls, 1973–1998. *Economics Letters*, 71(2):265–270, 2001.

[71] Nicholas G Rupp. The attributes of a costly recall: Evidence from the automotive industry. *Review of Industrial Organization*, 25(1):21–44, 2004.

[72] Nicholas G Rupp and Curtis R Taylor. Who initiates recalls and who cares? evidence from the automobile industry. *The Journal of Industrial Economics*, 50(2):123–149, 2002.

[73] Gabriel E Ryb, Cynthia Burch, Timothy Kerns, Patricia C Dischinger, and Shiu Ho. Crash test ratings and real-world frontal crash outcomes: a ciren study. *Journal of Trauma and Acute Care Surgery*, 68(5):1099–1105, 2010.

[74] François M Scherer. The welfare economics of product variety: an application to the ready-to-eat cereals industry. *The Journal of Industrial Economics*, pages 113–134, 1979.

[75] Aleksandr Shneyderman and Rodolfo Abella. The effects of the extended foreign

language programs on spanish-language proficiency and academic achievement in english. *Bilingual Research Journal*, 32(3):241–259, 2009.

[76] Michael Spence. Product differentiation and welfare. *The American Economic Review*, 66(2):407–414, 1976.

[77] Amanda L Sullivan. Disproportionality in special education identification and placement of english language learners. *Exceptional Children*, 77(3):317–334, 2011.

[78] Stephen J Trejo. Why do mexican americans earn low wages? *Journal of Political Economy*, 105(6):1235–1268, 1997.

[79] W Kip Viscusi. The value of risks to life and health. *Journal of economic literature*, 31(4):1912–1946, 1993.

[80] Brenda H Vrkljan and Dana Anaby. What vehicle features are considered important when buying an automobile? an examination of driver preferences by age and gender. *Journal of safety research*, 42(1):61–65, 2011.

[81] Clifford Winston and Fred Mannering. Consumer demand for automobile safety. *The American Economic Review*, 74(2):316–319, 1984.