Essays in Applied Microeconomics

by

Aviva Ronit Aron-Dine

B.A. Philosophy, Swarthmore College (2005)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Signature of Author ....

Department of Economics
May 4, 2012

Certified by ....

Amy Finkelstein
Professor of Economics
Thesis Supervisor

Certified by ....

David Autor
Professor of Economics
Thesis Supervisor

Accepted by ....

Michael Greenstone
3M Professor of Environmental Economics
Chairman, Departmental Committee on Graduate Studies

Essays in Applied Microeconomics

by

Aviva Ronit Aron-Dine

Submitted to the Department of Economics
on May 4. 2012. in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This dissertation consists of three chapters on topics in applied microeconomics. In the first chapter, I investigate whether voters are more likely to support additional spending on local public services when they perceive current service quality to be high. My empirical strategy exploits discontinuities in the Texas school ratings formula that create quasi-random variation in perceptions about school quality. I find that receiving an "exemplary" versus a "recognized" rating increases support for a school district's bond measures by about 10 percentage points. Voters respond to the level of a district's rating, not just to whether the district has improved or slipped. I develop and implement a test for whether these patterns of voter behavior lead to efficient outcomes; however, the results are inconclusive.

The second chapter, written jointly with Liran Einav, Amy Finkelstein, and Mark Cullen, investigates whether individuals exhibit forward looking behavior in their response to the non-linear pricing common in health insurance contracts. Our empirical strategy exploits the fact that employees who join an employer-provided health insurance plan later in the calendar year face the same initial price of medical care but a higher expected end-of-year price than employees who join the same plan earlier in the year. Our results reject the null of completely myopic behavior; medical utilization appears to respond to the future price, with a statistically significant elasticity of medical utilization with respect to the future price of -0.4 to -0.6. To try to quantify the extent of forward looking behavior, we develop a stylized dynamic model of individual behavior and calibrate it using our estimated behavioral response and additional data from the RAND Health Insurance Experiment. Our calibration suggests that the elasticity estimate may be substantially smaller than the one implied by fully forward-looking behavior, yet it is sufficiently high to have an economically significant effect on the response of annual medical utilization to a non-linear health insurance contract. Overall, our results point to the empirical importance of accounting for dynamic incentives in analyses of the impact of health insurance on medical utilization.

In the third chapter, I exploit a discontinuity in federal financial aid rules at age 24 to estimate the effect of financial aid on college enrollment, school choice, and persistence and degree completion rates. Undergraduate students who are not married and do not have children are classified as "dependent" or "independent" for purposes of federal financial aid based on whether they have turned 24 as of January 1 of the "award year." Independent students qualify for additional grant aid and are eligible to take out much larger federal loans. Using data from the National Postsecondary Student Aid Study and the Beginning Postsecondary Students Longitudinal Study, I show that average grant aid per student increases by about $1,100, or 55%, at age 24, while 12% of students take advantage of the higher federal loan limits. Estimates of the effects of additional aid on enrollment, persistence, and degree completion are inconclusive; while not statistically significant, they do not allow me to rule out sizable effects. I do find evidence of an increase in enrollment at for-profit colleges, concentrated among students whose parents are not college graduates.

JEL classification: D72, I13, I23

Thesis Superviser: Amy Finkelstein
Title: Professor of Economics

Thesis Superviser: David Autor
Title: Professor of Economics

Acknowledgments

I am grateful to my advisors, Amy Finkelstein and David Autor, for their feedback, advice, and encouragement. I feel lucky to have had the opportunity to work with them, and they have taught me a tremendous amount.

I have also benefited from many helpful conversations with other MIT faculty and with my fellow students. I especially want to thank David Chan, Gregory Leiserson, Christopher Palmer, Christopher Walters, and Tyler Williams for frequent and stimulating discussions.

Finally, I am grateful to my husband, Matthew Fiedler, my parents, Melanie Aron and Michael Dine, and my siblings, Jeremy and Shifrah Aron-Dine, for their ongoing love and support. Matt also provided extensive assistance and feedback at every stage of the research process, and this dissertation was immeasurably improved by his insights.

Contents

Chapter 1

How Does Perceived Service Quality Affect Voters' Willingness to Pay for Public Services?[1]

1.1 Introduction

In their capacity as voters, individuals regularly face decisions about their desired levels of taxes and spending. These decisions require them to assess whether or not additional public spending is worthwhile, a complicated inference problem with many potential inputs.

In this paper, I examine how voters' views about the quality of existing public services influence these decisions. Specifically, I test whether voters are more or less likely to support additional spending on local public services when they receive information telling them that current service quality is high. A priori, whether and how voters will respond to such information is ambiguous. One might expect that voters would be more willing to increase spending and taxes when they think the local government is doing a good job. Conversely, voters might reason that it is precisely when current quality is low that additional funding is needed. Or they might base their decisions on factors orthogonal to current quality.

In order to distinguish among these possibilities, I exploit quasi-random variation in perceptions of service quality created by Texas's school district ratings. Each year, Texas assigns school districts one of four ratings: exemplary, recognized, acceptable, or unacceptable, based on whether they meet cutoffs for standardized test pass rates and other indicators. As discussed below, these ratings are heavily publicized by school districts and in the press, and so it is likely that the ratings (and not

[1] I am grateful to Amy Finkelstein and David Autor for extensive feedback on this project. I also thank David Chan, Matthew Fiedler, Gregory Leiserson, Christopher Palmer, Henry Swift, Christopher Walters, Xiao Yu Wang, Tyler Williams, and participants in the MIT Public Finance Lunch for helpful comments and suggestions, Laura Slaughter and Aaron Bradley for assistance with data, and Miikka Rokkanen for help implementing the Imbens-Kalyanaraman bandwidth selection algorithm. Financial support from the National Science Foundation Graduate Fellowship is also gratefully acknowledged.

just the underlying formula inputs) affect individuals' perceptions of school quality.[2]

For districts close to a ratings cutoff, what rating the district receives depends on small, unpredictable fluctuations in test scores and other measures that should be outside the district's control and effectively random. This allows me to employ a regression discontinuity (RD) design strategy to obtain consistent estimates of the impact of the school district ratings on support for school bond measures. I interpret these estimates as measuring the effect of changes in perceived school quality on voting behavior.

My central finding is that improvements in perceived school quality increase support for school bond measures. In my preferred specification, receiving a rating of exemplary rather than recognized increases support for a district's bond measures by 10 percentage points. (Receiving a rating of recognized rather than acceptable has no statistically significant impact on support for school bond measures, although I cannot rule out an effect as large as 9 percentage points.) I find that voters respond to the level of their district's rating, not just to whether the rating is better or worse than the previous year's.

The main contribution of this paper is additional insight into how voters make decisions about the level of taxes and spending. The paper adds to the small but growing economics literature that uses quasi-experimental approaches to identify causal factors influencing voter preferences and decisions about these issues.[3] For example, Hoxby (2001) exploits differences among school finance equalization schemes to show that school districts impose higher taxes when more of the money is spent within the district, rather than elsewhere in the state, and Cabral and Hoxby (2011) and Finkelstein (2009) examine how cognitive limitations, specifically inattention and susceptibility to salience effects, influence opinions about taxes. My findings contribute to this literature by identifying another important input into voters' decisions about taxes and spending.

The results here also bear on two other literatures. First, my finding that the school district ratings affect voting behavior confirms my initial presumption that the ratings influence voters' perceptions. My results thus add yet another example to the growing list of settings in which individuals respond to discrete ratings cutoffs even though the underlying formula is publicly available.[4]

Second, the results have implications for the effect of introducing school ratings systems, as more than a dozen states have now done.[5] The existing literature on these systems focuses primarily on the penalties the systems impose on the worst-rated schools (see for example Figlio and Rouse (2006) and Reback (2008)). My results imply that introducing a rating system may also have important consequences for the funding capacity of schools on different sides of the ratings cutoffs. If funding is productive, then introducing a rating system may exacerbate the initial disparities

[2] Consistent with this, Figlio and Lucas (2004) find that similar Florida school ratings affect home values.

[3] A much larger political science literature relies primarily on survey evidence and focuses on identifying correlates of voter beliefs about taxes and spending. See for example Campbell (2009).

[4] For other examples, see Luca (2011) (Yelp), Luca and Smith (2011) (*U.S. News and World Report* college rankings), Pope (2009) (*U.S. News and World Report* hospital rankings), and Lacetera et al. (Forthcoming) (odometer readings).

[5] The No Child Left Behind law, which categorizes schools and districts as having made or failed to make "Annual Yearly Progress," is also a type of school rating system.

among schools. More broadly, the pattern of voter behavior I document here, in which voters are more willing to approve additional funding when current service quality is high, is a mechanism that will always tend to amplify initial differences in service quality (assuming the marginal product of funding is positive).

A natural question is what my results imply for the efficiency of public spending levels. Obviously, the large responses to miniscule differences in quality that result from voters reacting to the discrete ratings categories are probably not socially optimal.[6] More interesting to consider, however, are the efficiency implications of basing decisions about additional funding on current service quality.

Optimal policy requires that voters – whether by intent or merely in effect – set tax and spending levels based on the marginal productivity of public funds. Thus, the decision-making heuristic I have uncovered will produce efficient spending outcomes if and only if the total product of public funds (i.e. quality) is a good predictor of the marginal product. A standard, single-production function model of school quality would imply the opposite: lower-quality school districts would be expected to have higher returns to additional funding. On the other hand, it could also be the case that higher quality school districts operate on a different production function than lower quality school districts, one that is more steeply sloped for any reasonable level of inputs (perhaps due to more competent leadership).

In the penultimate section of the paper, I develop and implement a test for the correlation between initial quality and the marginal product of public funds in my setting. While the results are unfortunately inconclusive, I discuss ways the test could be improved to hopefully yield more definitive answers.

The rest of the paper proceeds as follows. Section 1.2 provides relevant background regarding the Texas ratings formula and Texas school bond elections, and Section 1.3 describes my data. Section 1.4 explains the estimation strategy and the procedure by which I construct the running variable. Section 1.5 presents the results, Section 1.6 examines potential threats to the validity of the research design, Section 1.7 discusses efficiency implications, and Section 1.8 concludes.

1.2 Institutional Background

1.2.1 The School District Ratings

The Texas school district ratings system (officially known as the "Texas Accountability Rating System") was first introduced for the 1993-1994 school year, with the first ratings issued in August, 1994. Since then, the Texas Education Agency (TEA) has published new ratings annually (except in 2003), placing school districts into one of four categories: exemplary, recognized, academically acceptable ("acceptable"), and academically unacceptable ("unacceptable").[7] The ratings are based

[6] It may be individually rational for voters to respond to the discrete categories, depending on the cognitive costs of acquiring information about the underlying formula inputs.

[7] In the first two years of the system, the bottom two ratings were instead labeled "accredited" and "accredited warned." The TEA also publishes ratings for individual schools; I focus on the district ratings because school bond

primarily on standardized test scores. but also depend on dropout and. in some years, attendance rates. (I provide a more detailed description of the ratings formula in Subsection 1.4.1 below.)

Several features of the Texas ratings system are useful for my purposes. First, despite changes in standards and criteria, the system has retained the same basic structure since it was first introduced in 1994, allowing me to pool many years of data. Second, there are no monetary rewards or penalties attached to any of the Texas ratings. This allows me to interpret voter responses to the ratings as entirely responses to perceived quality, rather than responses to additional or reduced state funding.[8] Third, while the TEA does publish all the ratings formula inputs, it does not publish any summary measure of district quality except the ratings. This makes it more likely that voters' perceptions will be influenced by the discrete ratings categories.[9]

Also important for my purposes, the Texas ratings receive substantial attention and publicity. Each of the Texas daily papers included in the Lexis-Nexis database consistently publish front-page stories about the ratings, often accompanied by editorials. The news stories and, where applicable, the editorials, invariably highlight local school districts' performance. In addition, Texas requires all school districts to post their ratings on their websites and to mail information about the ratings to parents along with students' report cards.

On average, over the 18 years in which the Texas ratings system has been in place, 8% of districts have been rated exemplary, 31% have been rated recognized, 59% have been rated acceptable, and 2% have been rated unacceptable. Figure 1.1 shows that these averages mask considerable variation over time. When the ratings system was first introduced, the overwhelming majority of districts were rated acceptable. Over the next seven years, due to some combination of actual improvement, learning about the system, and falling standards, most districts moved up to one of the top two ratings categories. In 2003, no ratings were issued, and in 2004, the TEA made substantial changes to the system, introducing new criteria and new standardized tests and raising standards. Under the new ("post-2004") rating system, many districts initially saw their ratings fall; ratings then rose again until new criteria were introduced for 2010-2011.

Table 1.1 gives the average transition probabilities for districts receiving a given rating (omitting the switch to the new ratings system in 2004). The table shows that most districts keep the same rating from year to year. However, the table also makes clear that it is not uncommon for districts to rise or fall one or even two ratings categories between years. For example, while about 55% of the districts that receive an exemplary rating in some year retain that rating the next year, 36% drop to recognized, and 8% drop to acceptable.

elections take place at the district level. If voters base their decisions on their individual school's rating, rather than the school district's rating, this will bias my results towards zero.

[8] The Texas ratings have significant direct consequences only for districts that are rated unacceptable, which must work with the TEA on an improvement plan. Districts that are rated exemplary are excused from compliance with certain minor regulations and reporting requirements.

[9] Florida is the other large state with a long-established and high-profile school ratings system. However, under the Florida system, top-rated schools receive financial rewards. In addition, Florida assigns each school district both a rating and a numerical score and publishes both.

1.2.2 Texas School Bond Elections

Texas requires school boards to obtain voter authorization for all bond issues. Legally, bonds can be issued for construction, maintenance, purchases of land or buildings, or certain investments in equipment or technology. While money is fungible, Section 1.7 presents evidence that school districts do in fact spend bond proceeds mostly or entirely on construction and maintenance. Thus, my measure of voters' willingness to pay for public services is more specifically a measure of their willingness to pay for capital improvements.

Bond refenda in Texas require a simple majority to pass. Referenda generally either include specific tax increases or authorize the school board to increase taxes as needed to pay the interest on the bond issue. Based on the arguments employed in favor and against, school bond measures appear to be widely understood as encompassing both spending and tax increases.

School boards have considerable latitude in scheduling school bond elections. Because of state requirements that voters be given advance notice of elections, the school board must make the decision to place a school bond measure on the ballot at least three months in advance of the vote. However, there is no requirement that school bond elections take place on the same day as statewide primary or general elections. In practice, about 25% of school bond elections coincide with a statewide election, and about 10% coincide with a high-profile election.[10]

1.3 Data and Summary Statistics

1.3.1 School District Data

I obtained data on Texas school districts from two sources: the TEA and the National Center for Education Statistics (NCES) Common Core data system. From the TEA, I have data for 1994-2011 for all school districts on all major elements of the school ratings system. (I defer a detailed discussion of these data elements to the next section, where I explain the ratings formula.) I also have data on a limited set of district characteristics, including total enrollment and enrollment by race and by free or reduced price lunch status.

From NCES, I have data on per-student total, capital, and instructional expenditures. Where applicable, I adjust these data for inflation using the CPI-U. I use these measures as covariates and also to examine the effect of passing a bond measure on spending in Section 1.7.[11]

[10] By a high-profile election, I mean a statewide election involving a presidential, senatorial, or gubernatorial election or a presidential primary. While outside the scope of this paper, the question of why school boards schedule bond votes for high versus low turnout elections is an intriguing one.

[11] The NCES data cover fiscal years, which run from July 1 to June 30 and thus coincide with school years. Because the NCES financial data are available only through fiscal year 2009, I use TEA financial data for fiscal years 2010 and 2011. I do not use the TEA data for earlier years because until fiscal year 2006 these data appear to include only a subset of expenditures.

1.3.2 Elections Data

I obtained data on school bond elections from the Municipal Advisory Council of Texas, an association of Texas municipal bond underwriters. The data include the date of the election, vote totals, the bond amount authorized, and the purpose of the bond issue. The data are available since 1996 and cover approximately 85% of all school bond elections held during the 1996-2011 period, with better coverage in later years.[12] In total, I have data on 2,070 school bond elections.

I match school district ratings data and characteristics to votes taking place in the year after the ratings are issued. Thus, for example, 1995-1996 ratings data and characteristics are matched to votes taking place between August 3, 1996 and August 3, 1997. For the remainder of this paper, references to elections data from a particular year refer to the year of the relevant ratings, not the year of the elections (i.e. 1996 elections are those held from August 3, 1996 to August 3, 1997).

1.3.3 Summary Statistics

Table 1.2 displays summary statistics for all school districts as well as broken down by ratings category and restricted to those districts with a school bond election in the following year. (I restrict the sample to 1996-2011 because, as noted above, I observe school bond elections only starting in 1996.) Over the course of the 15-year sample period, I observe 1,303 distinct school districts (about 1,200 in any given year), for a total of 17,198 district x year observations. Most districts are small: the median district enrolls only 842 students, although mean enrollment is almost 4,000. (There are currently about 5 million students enrolled in Texas public schools.) Texas is a relatively poor state: in the average district, 52% of students are eligible for free or reduced price lunch. On average, 9% of students are black, and 31% are Hispanic.

Higher-rated districts are higher-income and less diverse than lower-rated districts, and they spend more per student. Setting aside the small fraction of districts rated unacceptable, higher-rated districts are also smaller than lower-rated districts, probably because more of them are located in suburban areas where school districts are more fragmented. (All of these differences across ratings categories are statistically significant at the 1% level.)

Out of the full sample of 1,303 districts, 782 held at least one bond election between 1996 and 2011. For the most part, districts that hold elections look similar to those that do not. The one exception is that districts that hold elections are considerably larger, perhaps reflecting the fact that larger districts are more likely to have some facility in need of replacement, renovation, or major maintenance expenditures in any given year.

Table 1.3 provides additional election-related summary statistics broken down by ratings category. In any given year, an average of 13% of districts hold elections. The table shows that districts rated exemplary hold fewer bond elections than districts rated recognized or acceptable. In addition, their bond measures are smaller, and turnout is lower. These differences, which are

[12] The Texas Bond Review Board provides a complete list of all school bond votes but unfortunately without vote totals. There is no relationship between the ratings formula discontinuities and the probability that the Municipal Advisory Council vote data for a given election are missing.

significant at the 1% level, probably reflect the above-noted fact that higher-rated districts are generally smaller, together with the fact that smaller districts seem to have a lower propensity to hold school bond elections.

About three quarters of bond measures pass.[13] Conditional on holding a vote, districts rated exemplary are somewhat more likely to approve bond measures, but this difference is not statistically significant (at the 5% level).

1.4 Estimation Strategy

As noted in the introduction, the idea behind my empirical strategy is that, for districts sufficiently close to the ratings cutoffs, which side of the cutoff the district falls on is as good as randomly assigned. Hence, sharp discontinuities in voter preferences at the ratings cutoffs can be interpreted as the result of the ratings.

More formally, I implement an RD design with distance from the ratings cutoffs as the running variable and support for school bond measures as the outcome. Because my measure of distance from the ratings cutoffs does not perfectly predict actual ratings, I use a "fuzzy RD" approach, instrumenting for actual with predicted ratings. In this section, I explain how I construct my distance measures, provide a more precise description of the estimation strategy, and present results from the first stage regressions.

1.4.1 Quantifying Distance From the Cutoffs

The Ratings Formula

This subsection provides an overview of the Texas school ratings formula. Detailed explanations of each year's formula can be found on the TEA website.[14] (The TEA's summary of the 2010-2011 ratings system is reproduced in Figure 1.3 and conveys some sense of the complexity of the formula.)

The Texas school district ratings are based on districts' performance on a large set of indicators, including:

- Pass rates on state standardized tests in English language arts, writing, math, social studies, and science, evaluated for all students and for four subgroups: black, white, and Hispanic students and "economically disadvantaged" students (students eligible for free or reduced-price lunch). (25 indicators)
- High school completion rates for all students and the same four subgroups. (5 indicators)
- Seventh and eighth grade dropout rates for all students and the same four subgroups. (5 indicators)

[13] Interestingly, that fraction fell markedly beginning in 2008, at the start of the recession.

[14] http://ritter.tea.state.tx.us/perfreport/account/.

- Attendance rates for all students and the same four subgroups. (5 indicators)
- Pass rates on state tests for special education students. (1 indicator)
- Pass rates on state tests for English Language Learners. (1 indicator)
- The share of all students and of economically disadvantaged students achieving "commended performance" in English language arts and math. (4 indicators)

The specific set of indicators considered has varied by year, and in no year have all of the above indicators been used. However, in some years districts have been evaluated on as many as 40 separate criteria.

In order to achieve a given rating, a district must meet the standards for that rating for *all* applicable indicators. (Districts do not have to meet standards for student groups in which they have sufficiently few students, generally groups with fewer than 30 students or groups with fewer than 200 students that also comprise less than 10% of the district's student population.) The standardized test measures are at the heart of the rating system and are the binding indicators for the large majority of districts. 93% of districts rated recognized, 94% of districts rated acceptable, and 68% of districts rated unacceptable fail to achieve a higher rating because of one of the standardized test indicators, rather than any of the completion, dropout, or attendance rate indicators (i.e. they are eligible for the higher rating on the basis of the completion, dropout, and attendence rate indicators).[15]

Depending on the year, the indicator, and the rating involved, districts may have various options for meeting the ratings standards. First, they may meet the "absolute standard" for a given rating and indicator. For example, to achieve an exemplary rating in 2010, a district needed at least a 90% pass rate on all the standardized test indicators, high school completion rates of at least 95% for all groups, and seventh and eighth grade dropout rates of less than 0.2% for all groups.

In most years, for most indicators, meeting the absolute standard is the only option for achieving an exemplary rating. But for purposes of obtaining a recognized or acceptable rating, districts often have the option of instead meeting "required improvement." For example, the absolute standard for a recognized rating in 2010 was an 80% pass rate on all standardized tests. But if a district's standardized test pass rate was between 70% and 80%, it was considered to have met the standard if its rate of improvement since 2009 was sufficient to meet the 80% standard in two years. Notice that the required improvement option still produces a sharp cutoff for receiving a recognized rating. If a district's 2009 pass rate was 70%, for example, it meets the standard if and only if its 2010 pass rate is at least 75%.

[15] A noteworthy feature of the Texas rating system is that it puts most of the emphasis on the weakest students. The absolute standards for standardized test pass rates have ranged from 25% to 70% for a rating of acceptable and from 65% to 80% for a rating of recognized and have always been 90% for a rating of exemplary. Thus, except for districts at risk of being rated unacceptable, a district's rating is generally determined by the weakest third (or less) of students in the worst-performing student group. It is not clear whether voters understand that this is the dimension of quality the ratings are capturing.

In 2009 and 2010 only, districts could also meet standards by using adjusted standardized test results (the "Texas Projection Measure") in place of the raw scores.[16] Finally, in some years, districts could claim a specified number of exemptions, excluding their worst indicators from consideration.

1.4.2 Constructing a Distance Measure

As noted above, I obtained detailed data on nearly all elements of the ratings formula from the TEA. In particular, I have data for all districts for all years on standardized test pass rates and dropout, completion, and attendance rates for all student groups, allowing me to evaluate whether districts have met the absolute ratings standards and whether they have achieved required improvement. I also have data on the number of students in each student group (to evaluate minimum size requirements) and data on the Texas Projection Measure and on allowable exemptions.[17]

In order to implement my proposed RD design, I need to turn the 46 distinct ratings criteria and the even larger number of separate standards into measures of districts' distance from each ratings cutoff.[18] My approach is as follows. I first compute each district's distance from each of the ratings cutoffs for each indicator, basing my calculation on the most lenient available option for meeting the standard.[19] Next, I apply the relevant exemptions, dropping those indicators for which the district is furthest from meeting the standard. I use the minimum of the remaining distances as my measure of distance from the ratings cutoff. Districts should be eligible for the higher rating if their distance from the cutoff is greater than or equal to 0 and ineligible if their distance is less than 0. (I normalize the final distance measures by dividing them by their standard deviations, to put them into interpretable units.)

To make this clearer, consider as an example a district that needs to meet only the English language arts, writing, and math pass rate standards for all students and is not eligible for any exemptions. The absolute standard for an exemplary rating is 90, and the absolute standard for a recognized rating is 80. The district's scores are 90, 82, and 75; last year, its scores were 70, 70, and 70.

To achieve a rating of exemplary, the district must meet the absolute standard for all three indicators. Thus, its distances from meeting the standards are 0, -8, and -15. Taking the minimum, the district's distance from the exemplary cutoff is -15; it is not eligible for a rating of exemplary.

To achieve a rating of recognized, the district can either meet the absolute standard or achieve required improvement, which, given its prior-year scores, means achieving a pass rate of 75 or

[16] Since the Texas Projection Measure adjusts the raw scores according to a predetermined formula, there is no reason to think that the existence of this option undermines the RD design.

[17] One might worry that districts could manipulate the minimum size requirements or student classification more generally. I address this concern in Section 1.6 below.

[18] I considered using Papay et al.'s (2011) approach to implementing an RD design with multiple assignment cutoffs, but concluded that this approach was neither feasible (with 46 assignment criteria) nor particularly informative in this setting. The key advantage of Papay et al.'s proposed method is that it allows one to estimate how treatment effects vary depending on which criteria bind, which is not a question of particular interest here.

[19] Using this approach, the standardized test measures end up on a very different scale than the dropout rate measures, with standard deviations on the order of 10 versus 2. In an attempt to increase power, I rescale each measure by dividing by its standard deviation. Omitting this normalization has essentially no impact on the results.

above on all three tests. Thus, achieving required improvement is the more lenient option, and the district's distance from meeting the standards are 15, 7, and 0. Taking the minimum, the district's distance from the recognized cutoff is 0; it just barely qualifies for a rating of recognized.

Note that, because different options are available for meeting the standards for different ratings, as well as because different criteria may be binding for different ratings cutoffs, the distance from the exemplary cutoff will not be a linear transformation of the distance from the recognized cutoff, which will not be a linear transformation of the distance from the acceptable cutoff. (Figure 1.2 illustrates this graphically.) For this reason, I in fact have three running variables – the distances from each of the cutoffs – not one.

There are two reasons that these distance measures will not perfectly predict actual ratings. First, my data do not allow me to implement certain minor elements of the ratings formula.[20] Second, districts have access to an appeals process: a district dissatisfied with its initial rating can file an appeal with the TEA. Appeals are supposed to be based primarily on data and computation errors, but the TEA has considerable discretion to grant appeals for other reasons. Since these appeals decisions are not made based on predetermined objective criteria, I cannot (and would not want to) incorporate them into my computations of distance from the ratings cutoffs.

1.4.3 Estimating Equation

I implement a "fuzzy RD" design where distance from the ratings cutoff is the running variable and predicted rating – specifically, an indicator for whether the distance measures described above are greater than or equal to 0 – instruments for actual rating. My basic specification uses local linear regression with a uniform kernel, and so the estimating equations are:

$$YesVoteShare_{d,t+1,e} = \beta_0 + \mathbf{1}(D_{dt} >= 0)\beta_1 + D_{dt}\beta_2 + \mathbf{1}(D_{dt} >= 0) * D_{dt}\beta_3 + \epsilon_{d,t+1,e} \qquad (RF)$$

$$Rating_{dte} = \delta_0 + \mathbf{1}(D_{dt} >= 0)\delta_1 + D_{dt}\delta_2 + \mathbf{1}(D_{dt} >= 0) * D_{dt}\delta 3 + \mu_{dte} \qquad (FS)$$

where D denotes distance from the ratings cutoff, $YesVoteShare$ is the share of voters voting in favor of the school bond measure, d indexes districts, t indexes years, and e indexes elections (some districts hold multiple elections in a given year). Year here refers to "rating year;" that is, years run from when one set of ratings is published to whenever the next set of ratings is published (generally August to August).

I estimate separate regressions for each of the ratings cutoffs. My preferred specification uses a bandwidth of 0.5 standard deviations of the distance measure, which is approximately the optimal bandwidth implied by the Imbens and Kalyanaraman (2009) algorithm and also seems reasonable based on visual inspection. As shown in Subsection 1.5.2 below, my results are robust to using a variety of alternative bandwidths and to using quadratic or cubic rather than local linear regression. Standard errors are clustered at the district x year level.

[20] In particular, ratings also depend on data quality criteria, and I do not have access to the state's data quality measures.

1.4.4 First Stage Results

Figures 1.4 and 1.5 graphically depict the first stage, or, equivalently, my success at matching the actual district ratings. The X-axis plots distance from the relevant ratings cutoffs, computed as described in Subsection 1.4.2. The Y-axis plots actual assigned ratings, with these ratings converted into numbers (0 = unacceptable, 1 = acceptable, 2 = recognized, and 3 = exemplary). The datapoints represent means for 0.1 standard deviation cells of the distance measure. Figure 1.4 plots the results for the primary estimation sample, districts with school bond elections, while Figure 1.5 plots the results for all district x year observations.

Predicted ratings match actual ratings 94% of the time in the vote sample and 90% of the time in the full sample.[21] In the graphs plotting distance from the exemplary and recognized cutoffs, there is a sharp break at 0, with the average rating increasing by almost a full rank. In contrast, there is at most a small increase in the average rating where distance from the acceptable cutoff equals 0. This is largely because most of the districts with distances less than 0, which should – according to the formula – be rated unacceptable, are instead rated acceptable. Apparently, the TEA is quite receptive to appeals from districts at risk of an unacceptable rating.

Table 1.4 shows the regression results for both samples. I have again converted the ratings into numbers, and so the coefficient estimates can be interpreted as fractions of a rating. Consistent with the graphical evidence, the first stage at the exemplary and recognized cutoffs is strong and highly statistically significant in both samples. The average rating increases by almost one full rating at the exemplary cutoff and by about four fifths of a rating at the recognized cutoff.

In contrast, there is no statistically significant first stage at the acceptable cutoff in the elections sample and only a small (three tenths of a rating) first stage in the full sample. Because most districts that would be rated unacceptable under the formula are instead rated acceptable, meeting the formula standards for an acceptable rating has little or no effect on a district's actual rating. This means that there is no quasi-experiment for perceived service quality around the acceptable cutoff. In the remainder of the paper, therefore, I restrict my analysis to the exemplary and recognized cutoffs.

1.5 Results

1.5.1 Main Findings

Figure 1.6 depicts the main reduced form results. The figure plots the average vote share in favor of the school bond measure ("yes-vote share") for 0.1 standard deviation cells of the distance measure. As the figure shows, yes-vote share is basically flat below the exemplary cutoff, averaging around 60%, and then jumps sharply to about 70% at the cutoff. In contrast, yes-vote share increases smoothly with distance from the recognized cutoff, but there is no discontinuous break.

[21] The match rate is higher in the vote sample because a disproportionate fraction of votes occur in later years, where I am more succesful at matching the ratings.

Table 1.5 presents estimates from two-stage least squares regressions of the effect of a higher rating on yes-vote share. The 2SLS regressions scale up the reduced form effects depicted in the figure to reflect the fact that the distance measures do not perfectly predict actual ratings. As shown in Table 1.4, the scaling factor is about 1/0.95 at the exemplary cutoff and about 1/0.85 at the recognized cutoff.

Consistent with the graph, the regression results indicate that receiving an exemplary rather than a recognized rating leads to a roughly 10 percentage point increase in yes-vote share ($p = 0.01$). The point estimate is essentially unchanged when I add year dummies and district characteristics to the regressions, although the standard errors fall modestly. While the point estimates for the effect of receiving a recognized versus an acceptable rating (shown in columns 4-6 of Table 1.5) are positive, they are considerably smaller than the estimated effects of receiving an exemplary rating and not statistically significant.[22]

How large is a 10 percentage point increase in yes-vote share? As one way of putting this estimate in context, the average within-district standard deviation of yes-vote share (among districts that hold more than one school bond election during the sample period) is 11 percentage points. Thus, the impact of receiving an exemplary versus a recognized rating is sizable relative to typical fluctuations in a given district's election outcomes.

To really interpret the magnitude of the estimated response, however, one would want to know how much the ratings influence voters' perceptions of quality. As discussed in Subsection 1.2.1 above, the Texas ratings receive enough publicity that one would expect them to have at least some effect on perceptions; moreover, the existence of a discontinuity in yes-vote share at the exemplary cutoff confirms that the ratings influence voters' views to some extent. Ideally, though, I would be able to directly estimate the true first stage: the effect of the ratings on perceived school quality. I considered whether it might be possible to do this using data from a website such as greatschools.org where parents rate their children's schools. Unfortunately, none of these websites seem to be sufficiently popular in Texas. Thus, I would probably need to field my own survey, asking potential voters what they think of the local schools.

For now, as a simple benchmark, I impose the assumption that voters think their school district is as good as the median district with the same rating. For example, in a year in which 40% of districts are rated recognized and 10% are rated exemplary, I assume that voters in districts rated recognized think their districts are at the 70^{th} percentile of district quality, and voters in districts rated exemplary think their districts are at the 95^{th} percentile of district quality. Under these assumptions, I find that voters would rank exemplary districts an average of 25 percentiles above recognized districts in the district quality distribution. This implies that each percentile increase

[22] The estimate for the effect of receiving an exemplary rating remains highly statistically significant ($p = 0.02$) even with a correction for multiple inference. If I instead pool the cutoffs, the point estimate (from the specification including year effects and district characteristics, which has the smallest standard errors) is that a higher rating increases yes-vote share by 5 percentage points, with a p-value of 0.055. This finding is a mechanical result of the fact that the point estimate at the exemplary cutoff is large and positive, while the point estimate at the recognized cutoff is small but positive. Because the point estimates at the two cutoffs are so different, I regard the results from the separate regressions as more meaningful.

in perceived quality is worth a 0.4 percentage point increase in yes-vote share. If the actual impact of the ratings on voters' perceptions is less (more) than under my benchmark assumption, then the implied effect of perceived quality on voting behavior is larger (smaller).[23]

It is not clear what to make of the fact that voters appear to respond strongly to an exemplary but not a recognized rating. The estimates are fairly imprecise, and I cannot rule out increases in yes-vote share at the recognized cutoff as large as 9 percentage points. Thus, it is possible that the responses at the two ratings cutoffs are not actually that different (or even different at all), and I simply lack the statistical power to detect the response at the recognized cutoff.

It is also possible that voters really do respond differently at the two cutoffs. Perhaps voters care more about whether their district is ranked at the very top of all districts than about whether it is ranked in the upper or lower middle of the distribution, or perhaps voters in higher-performing districts pay more attention to the school ratings than voters in lower-performing districts.

Both of the above hypotheses imply that receiving a recognized versus an acceptable rating should have more of an impact on yes-vote share in years when the ratings criteria are more stringent and fewer districts rank as high as recognized. As Figure 1.1 shows, there is considerable variation across years in the share of districts rated recognized or above; the recognized category is most selective from 2004-2008. Intriguingly, when I split the sample into the 2004-2008 period and all other years, the estimated effect on yes-vote share is 5 percentage points ($p = 0.16$) in the years where the recognized category is most selective, versus 2 percentage points ($p = 0.54$) in the other years. While the difference between these estimates is not statistically significant, the results are consistent with the notion that information about district quality may have a larger impact on voting behavior in higher quality districts.

1.5.2 Robustness and Specification Checks

Table 1.6 shows that my main results are highly robust to alternative regression specifications. Allowing bandwidth to range from 0.3 to 1.0 standard deviations of the distance measure, the point estimates for the effect of an exemplary rating on yes-vote share range from 8 to 13 percentage points, with p-values ranging from 0.01 to 0.03. If I instead estimate the effect of the higher rating using a quadratic or cubic polynomial (including all bond elections in the regressions and allowing the coefficients on the polynomial to differ above and below the cutoff), the point estimates are somewhat larger (16 to 18 percentage points), with p-values of 0.002. The estimates of the effect of a recognized versus acceptable rating are also robust to alternative specifications, in that the point estimates remain small and statistically insignificant.

Table 1.7 changes the dependent variable, estimating the effect of a higher rating on the probability that a district's bond measure passes. While the point estimate indicates that receiving an exemplary versus a recognized rating improves the odds of passage by 14 percentage points, the

[23] At first blush, it seems unlikely that the ratings would shift voters' perceptions this much, since voters presumably have other sources of information about school district quality besides the ratings. On the other hand, my results may be driven by low information voters for whom the ratings are a or even the main source of information about school quality.

standard error is large, and the p-value is 0.14. Naturally, substituting a binary for a continuous outcome measure provides less information and less statistical power. Moreover, it is not clear that one should expect large effects on yes-vote share to translate into large effects on passage rates in this setting. As noted above, mean yes-vote share just below the exemplary cutoff is about 60%. Given that bond measures in Texas need only a simple majority to pass, even large increases in yes-vote share may have little effect on the fraction of bond measures approved.

The estimated effect of receiving a recognized versus an acceptable rating on passage rates is small (less than 1 percentage point) and not statistically significant.

1.5.3 Extensions

Levels Versus Changes

As explained in Subsection 1.2.1 above, the TEA issues new school district ratings each year. As shown in Table 1.1, the majority of districts retain the same rating from year to year, while some move up in the ratings, and a smaller fraction move down. One might expect that, in districts where the ratings do not change from year to year, the ratings announcements would receive less attention, and the ratings would cease to affect behavior. On the other hand, it could also be the case that voters attach less weight to the ratings in districts where the ratings have changed. Suppose that voters are fully rational in how they process the information provided by the ratings (but do not look at the underlying formula inputs, because doing so is costly). In that case, voters would take into account that a district where the rating has changed is probably closer to the rating boundary than a district where the rating has stayed the same. For example, consider four districts. District A is rated exemplary in years 1 and 2, district B is rated recognized in years 1 and 2, district C is rated recognized and then exemplary, and district D is rated exemplary and then recognized. When going to the polls in year 2, rational voters should perceive a larger quality differential between districts A and B than between districts C and D.

The first panel of Figure 1.7 and column 2 of Table 1.8 restrict the sample to districts that have received the same rating for at least two years, while the second panel of the figure and column 3 of the table show results for districts whose most recent rating differs from their prior-year rating. Consistent with rational information processing, voters seem to respond more strongly to the ratings in districts where the ratings remain constant from year to year. The point estimate for the effect of an exemplary versus a recognized rating among districts whose ratings did not change is essentially the same as the full sample point estimate, while the point estimate for the districts whose ratings did change is considerably smaller. However, the difference between the two estimates is not statistically significant. Thus, the evidence that voters respond less to the ratings where the ratings have changed is only suggestive. What is clear is that voters respond to the level of their district's rating, not just the change in the rating. Equivalently, they respond to information about quality levels, not just information about changes in quality.

Panels C and D of Figure 1.7 examine whether voters respond differently to increases versus decreases in their district's rating. The sample in Panel C is restricted to districts that received a rating of recognized or acceptable the previous year. Thus, in this graph, districts that are to the right of the cutoff have improved their rating relative to last year, while districts to the left of the cutoff have mostly held steady. Meanwhile, the sample in Panel D is restricted to districts that received a rating of exemplary the previous year. Thus, in this graph, districts that are to the left of the cutoff have slipped a rating since last year, while districts to the right have held steady. One can therefore interpret the discontinuity at the cutoff in Panel C as voters "rewarding" improvement and the discontinuity at the cutoff in Panel D as voters "punishing" slippage.

Looking at the graphs, it seems more clear that voters are rewarding improvement than that they are punishing slippage. Unfortunately, however, there is a large difference in sample size between the two graphs, since districts are much more likely to improve than to slip. Columns 4 and 5 of Table 1.8 show that the estimated increase in yes-vote share due to improvement is about equal to the estimated decrease in yes-vote share due to slippage (both are about 10 percentage points), but only the former is statistically significant.

Rational voters would distinguish slippage in the ratings due to changes in a district's performance from slippage due to changes in the ratings system. Thus, rational voters would not be expected to punish districts for the drop in the ratings that occurred for the large majority of districts following the switch to the new ratings system in 2004. Interestingly, when I omit 2004 from the regressions in Table 1.8, the point estimate in the slippage column increases from 0.11 to 0.15. However, the estimate is still not statistically significant.

Figure 1.8 and Table 1.8 Panel B provide the corresponding breakdown for the effects of a recognized versus acceptable rating. In this case, none of the estimates are statistically significant or strikingly different from one another.

Timing and Types of Elections

As a plausibility check on the main results, I examine how the effect of receiving a higher rating on yes-vote share varies with the length of time between when the ratings are issued and when the election is held. In most years, the TEA publishes the ratings around August 1, at which point, as discussed in Subsection 1.2.1 above, they receive considerable media attention. Districts then have the opportunity to appeal their ratings; appeals decisions are issued in October and typically result in another flurry of local news stories.

School bond elections can be held at any time but are most commonly scheduled for September, November, February, or May. I divide each year in my sample into four quarters beginning when the ratings are issued in August; each quarter includes one major election date. I then estimate the baseline regressions separately for the four quarters.

The results are shown in Table 1.9. Reassuringly, the effect of receiving an exemplary rating on yes-vote share is larger for elections held sooner after the ratings are issued. Interestingly, this is not the case for the effect of receiving a recognized rating, perhaps suggesting that the effect at the recognized cutoff is a true null.

Also interesting to consider is how the effect of receiving a higher rating on yes-vote share varies with voter turnout rates. While I do not directly observe turnout rates, it seems reasonable to assume that turnout will be highest for school bond elections that coincide with "high-profile" state or national elections (defined as in Subsection 1.2.2 above as presidential, gubernatorial, and senatorial elections and presidential primaries) and higher for school bond elections that coincide with statewide elections than for those that do not.[24] One might expect the voters who turn out for lower-profile school bond elections to be those with strong opinions about the bond measure, who might be less swayed by the ratings. On the other hand, it might also be that the voters who care enough to turn out for elections where only a school bond measure is on the ballot also pay more attention to the school ratings.

The results, shown in Table 1.10, indicate that the large effect of receiving an exemplary rating on yes-vote share is driven entirely by school bond votes that do not coincide with statewide elections. Since only a small number of elections in districts close to the exemplary cutoff coincide with statewide and especially with high-profile elections, the estimates for these samples are imprecise, and I cannot rule out the possibility that the ratings affect voting behavior in these elections as well. What is clear, however, is that voters in low-turnout, off-cycle school bond elections make decisions partly based on the ratings.

Mechanisms: Effects on Turnout

There are two possible mechanisms by which the ratings could affect yes-vote share. They could cause individual voters to change their minds, or they could change the mix of voters coming to the polls. For example, individuals might be more motivated to vote in the school bond election if they feel more positively about the local schools.

Figure 1.9 and Table 1.11 attempt to look at this by estimating the effect of the ratings on log turnout. While the results are somewhat imprecise, there is no evidence that turnout increases at either of the ratings cutoffs. This suggests that the more likely mechanism behind the results is that some voters actually vote differently when the district is rated exemplary versus recognized.

1.6 Threats to Validity

The validity of any regression discontinuity design depends on the assumption that small fluctuations in the running variable around the cutoff are as good as random. It is not necessarily a problem if subjects can manipulate the running variable; for example, in my setting, it would not necessarily be a problem if some districts cheat on state tests. What would be a problem would be if certain districts can manipulate their scores precisely enough to systematically move themselves

[24] An alternative would be to define high turnout elections based on actual turnout levels. However, since I do not observe turnout rates, only the number of voters, and since large districts may schedule elections differently than smaller districts, this approach could produce arbitrary sample splits. I considered proxying for turnout rates by dividing the number of voters by the number of students enrolled in the school district but was concerned that there might be large differences across districts in the ratio of eligible voters to school-age children.

from just below to just above the ratings cutoffs. Because successful manipulators might differ from less successful manipulators in their underlying levels of support for school bond measures, such manipulation could bias the estimates.

There are two standard approaches to testing for problematic manipulation of the running variable. The first involves testing for a discontinuity in the density of observations at the cutoff; manipulation will produce "bunching" on the more favorable side of the discontinuity (McCrary, 2008). The second involves testing for discontinuities in covariates around the cutoff; manipulation will produce such discontinuities if manipulators are systematically different from non-manipulators. I implement both of these tests. To increase power (and thereby increase the likelihood that I will identify problems), I implement the tests in the full sample of all district x year observations, rather than in the smaller sample of districts with bond elections.[25] I also consider another possible threat to validity: that the ratings may influence school boards' decisions about whether to schedule bond elections.

1.6.1 Density Analysis

As explained in Subsection 1.4.1 above, in practice, the Texas school district ratings depend primarily on standardized test pass rates. For more than 90% of districts rated recognized or acceptable, the binding constraint preventing them from receiving a higher rating is one of the test score measures. I therefore focus on testing for manipulation of test scores.[26]

In general, one would not expect school districts to be able to manipulate standardized test pass rates precisely enough to cause any problems for the validity of the RD, particularly given that tests are scored centrally, not by the individual schools or districts. The problem is that ratings outcomes depend not just on test performance but also on which students' tests are counted and which students are counted as belonging to which student groups. In the early years of the Texas ratings system, school districts had an opportunity to "correct" their enrollment and demographic files after receiving the standardized test results. This opportunity was eliminated in the post-2004 system. In addition, the post-2004 ratings system is much more complex and nonlinear than the old system, making it more difficult for districts to determine which few students' scores are the obstacle to a higher rating.

Figure 1.10 plots the density of test outcomes in the full, pre-2004, and post-2004 samples. Datapoints in this figure are numbers of standardized test indicators (rather than numbers of

[25] When I implement the tests in the sample of districts with bond elections, I find no evidence of manipulation, but the estimates are also much less precise.

[26] I did attempt to test for manipulation of the other indicators (dropout, completion, and attendance rates), but it was difficult to get any traction since so many districts were "heaped" at dropout rates of 0 or completion rates of 100%. Table 1.15 shows that the estimated effect of an exemplary rating on yes-vote share is virtually unchanged if I construct the distance measure using only the standardized test indicators, and so manipulation of the other indicators cannot be driving this result. When I exclude the non-test score indicators from the construction of the running variable, I find a statistically signficant 5 percentage point increase in yes-vote share at the recognized cutoff. However, this is the only specification in which I find a statistically significant response to receiving a recognized versus an acceptable rating, and the result would not survive a multiple inference correction.

districts) at a given distance from the relevant cutoff. For example, if the cutoff for recognized is a pass rate of 75% and a district has three student subgroups with pass rates of 72% in math, then the district contributes three standardized test indicators at a distance of -3 percentage points from the recognized cutoff. Manipulation would be expected to result in too many test indicators just at the cutoff and too few just below it.

From the figure, it does not appear that there is any discontinuity in the density at the exemplary cutoff. But there is clear evidence of bunching above the recognized cutoff. The remaining panels of the figure show that this bunching is driven entirely by the earlier period; in the post-2004 sample, the density around the recognized cutoff is quite smooth.

Table 1.12 presents the corresponding regression results. The regression specification is the same as for the primary results, except that the outcome variable is now the number of standardized test indicators at a given distance from the cutoff (the bandwidth is 5 percentage points). The regressions provide no evidence of a discontinuity in the density at the exemplary cutoff or at either cutoff in the post-2004 sample, but strong evidence of a discontinuity at the recognized cutoff in the earlier sample.

1.6.2 Covariate Balance

Figures 1.11 and 1.12 and Tables 1.13 and 1.14 examine whether the apparent manipulation of test scores gives rise to any imbalances in covariates around the cutoffs.[27] The figures present results for the full sample, but the tables also show results for the pre-2004 and post-2004 samples.

I examine the share of a district's students that are black, Hispanic, or eligible for free or reduced price lunch as well as total enrollment, prior-year test pass rates, and per-student spending.[28] There are no statistically significant discontinuities, even in the pre-2004 sample. While some of these results are noisy, the results for student characteristics and prior-year test scores are quite precise, especially at the recognized cutoff, which is the cutoff at which there is evidence of manipulation. For example, I can rule out a discontinuity of more than 0.019 percentage points in the share of students eligible for free or reduced price lunch and a discontinuity of more than 0.6 percentage points in prior-year pass rates.

1.6.3 Discussion

Despite the absence of any discontinuities in observable covariates, the manipulation documented in Subsection 1.6.1 is troubling, since, in principle, differences between manipulators and non-manipulators could be driving my results. To address this concern, Table 1.15 shows results for the effect of a higher rating on yes-vote share restricting the sample to the post-2004 period, where

[27] The data in the figures are purged of year effects, primarily because the per-student spending series is otherwise extremely noisy.

[28] The prior-year test pass rate variable is computed as the average (across subjects) share of students in the district passing state tests. The ideal covariate to examine would be prior election yes-vote share. Since most districts in my dataset hold at most one election during the sample period, that measure is not available.

there is no evidence of (and should be no opportunity for) manipulation. As shown in Panel A column 2. the estimated effect of an exemplary rating on yes-vote share in the post-2004 period is highly statistically significant ($p = 0.01$) and is actually larger than the full-sample estimate. though not statistically distinguishable.

Given that the covariate balance tests provide no evidence that the pre-2004 manipulation leads to important differences between districts just above and just below the cutoff. and given that there is no evidence of manipulation at the exemplary cutoff even in the pre-2004 subsample. I continue to prefer the full sample results, which are considerably more precise. The fact that these results are robust to excluding the early years is highly reassuring, in that it indicates that manipulation cannot be the driving force behind the main findings.

1.6.4 Endogeneity of Elections

In addition to the standard concerns about the validity of the RD, a concern specific to my setting is that the availability of the outcome variable may be endogenous to the discontinuity. That is, I only observe yes-vote share when a district holds an election, and school boards might be more or less likely to schedule bond elections depending on the district's rating.

Figure 1.13 and Table 1.16 attempt to test this concern directly by examining the effect of a higher rating on a district's propensity to hold elections. As indicated in Table 1.3, on average, higher-rated districts are less likely to hold elections. However, there is no evidence of a discontinuity in the propensity to hold an election at either the exemplary or the recognized cutoff.[29]

While the standard errors on these results are large, two other considerations should be reassuring. First, if school boards are more likely to place bond measures on the ballot in years where the district receives a higher rating, this is presumably because they think voters are more likely to approve the measures in those years. Thus, assuming school boards have an accurate understanding of the electorate, endogenous scheduling of elections could bias the magnitude of my results but should not lead me to inappropriately reject the null.

Second, as explained above, school bond elections in Texas must be scheduled at least three months in advance, and so elections held within three months of the publication of the ratings should already have been scheduled when the ratings were issued. While it is possible that school boards anticipate the district's rating and act accordingly, it is still somewhat reassuring that, as shown in Table 1.9, the effect of receiving an exemplary versus a recognized rating is strongest for elections held within three months of the date the ratings are issued.

1.7 Implications for the Efficiency of Spending Levels

The results discussed above enhance our positive understanding of voter behavior, as well as of the effects of school ratings systems. But they do not answer the normative question, posed in the

[29] I also test for and find no evidence of a discontinuity in the size of the bond issues school boards place on the ballot.

introduction. of whether these patterns of voter behavior result in efficient spending levels.

For school bond election outcomes to be efficient, bond measures must pass if and only if the marginal product of new capital spending exceeds the marginal cost of public funds. It seems unlikely that the school district ratings provide voters with new information about the marginal cost of public funds. but more plausible that they might provide information about the marginal product. Put differently, it is possible that the total product of public funds (i.e. school district quality) predicts the marginal product in this setting. One can easily imagine reasons why this might be the case. For example, more competent school boards might both achieve higher ratings and allocate the district's capital budget more productively. At the same time, it also seems plausible that the value of new facility investments might be greatest in the worst-performing school districts.

To evaluate whether the voting behavior I have documented leads to efficient spending levels, I need to determine the correlation between quality (as measured by the ratings formula) and the marginal product of capital spending. Below, I outline a procedure to estimate this correlation and present the – unfortunately inconclusive – results. I then discuss how the test could be improved to perhaps give more definitive answers.

1.7.1 The Test

In order to test for the correlation between school district quality and the marginal product of capital spending, I first need a way to estimate the treatment effect of additional capital spending. To do this, I take advantage of the fact that, in sufficiently close elections, whether a bond measure passes (i.e. whether it receives just over or just under 50% of the vote) is effectively random. This allows me to again employ a "fuzzy RD" design where yes-vote share is now the running variable and passing a bond measure instruments for capital spending. (The same basic approach is used by Cellini et al. (2010) to estimate the effect of California school bond measures on housing values and test scores.) I then stratify the results on "initial" distance from the exemplary cutoff (distance from the cutoff in the year before the bond measure) to estimate treatment effects for districts at different levels of initial quality. The relationship between these treatment effects gives me the correlation between the marginal productivity of funds and initial quality.

The estimating equations are:

$$TestPassRate_{dt+4} = \beta_0 + \mathbf{1}(V_{dte} > 0.5)\beta_1 + V_{dte}\beta_2 + \mathbf{1}(V_{dte} > 0.5) * V_{dt}\beta_3 + \epsilon_{d,t+4,e} \qquad (RF)$$

$$\sum_{k=0}^{4} Capital_{dt+k} = \delta_0 + \mathbf{1}(V_{dte} > 0.5)\delta 1 + V_{dte}\delta_3 + \mathbf{1}(V_{dte} > 0.5) * V_{dte}\delta_3 + \mu_{d,t+4,e} \qquad (FS)$$

where V is the vote share in favor of the school bond measure, $TestPassRate$ is the average (across subjects) share of students in the district passing state tests, and $Capital$ is per-student capital expenditures. d indexes districts, t indexes fiscal/school years, and e indexes elections. (Fiscal and school years coincide, since the fiscal year runs from July 1 to June 30.) For reasons explained in Subsection 1.7.2 below, I estimate the effects on test scores four years after the election. Presumably, test scores depend on the total additional capital spending that results from the bond measure. rather than just on capital spending in the year of the test. and so the endogenous variable is total construction spending in the year of the vote plus the subsequent four years. I use a bandwidth of 10 percentage points of vote share (the results are generally insensitive to bandwidth selection). All regressions include year dummies and control for pre-election district characteristics. including spending and test scores. to enhance precision. Standard errors are clustered at the district x year level.

To determine the correlation between the treatment effect and quality, I rerun the above regressions splitting the sample approximately in half based on distance from the exemplary ratings cutoff in the year before the election.[30] The quantity of interest is the difference between the two-stage least squares coefficients on construction spending estimated in the high versus low quality samples.

Note that the results I obtain using this strategy are local average treatment effects for districts with close bond elections. If the correlation between initial quality and the marginal productivity of new capital spending is different for these districts than for other districts, then this strategy could give the wrong answer to the question of whether voting based on initial quality produces efficient outcomes on average. Table 1.17 compares districts with close elections (defined as those where a measure passed or failed by 5 percentage points or less) to other districts with votes. For the most part, the two samples look fairly similar on observable measures, although the districts with close elections are larger and spend less per student at baseline.

1.7.2 First Stage Results

One possible concern about the above research design (discussed extensively in Cellini et al. (2010)) is that the outcome of close elections may not actually have much effect on capital spending, because school boards in districts where the school bond measure goes down to defeat may just keep putting measures on the ballot until one eventually passes. There is certainly some evidence of this phenomenon in my sample. Districts where a bond measure just barely fails are almost 50 percentage points more likely than districts where a measure just barely passes to hold a second election within four years.[31]

Nonetheless. Figure 1.14 and Table 1.18 show that passing a bond measure has a strong and statistically significant effect on capital spending. at least at lags up to four years. In the first panel of the graph and column of the table. the dependent variable is per-student capital spending in the year of the election. Each additional panel and column add an additional year of capital

[30] This approach splits the sample at one standard deviation below the exemplary cutoff. Given that I find that voters' response to the ratings is concentrated around the exemplary cutoff, it would be better if I could split the sample at that cutoff. Unfortunately. there are not enough close elections in districts above the exemplary cutoff to make that feasible. Ideally. of course, I would be able to trace out the relationship between initial quality and the treatment effect over the entire distribution of quality.

[31] This estimate comes from running the above specification with a dummy for holding a second vote on the left hand side.

expenditures to the dependent variable (that is. the dependent variable is cumulative).

Reassuringly. in the year of the vote. there is essentially no difference in capital spending between districts where a bond measure just barely fails and districts where a measure just barely passes. In each of the next two years. passing a bond measure leads to about a $1.500 per student increase in capital expenditures (compared to a base year mean of about $900). and in the third year it leads to an increase of about $1.000. By the end of three years. districts where a measure just barely passed have spent nearly $4.000. or about 70%. more per student than districts where a measure just barely failed. The estimated difference in cumulative spending falls a bit by the end of year four, presumably because many districts where the inital bond measure failed have by then managed to approve a subsesquent measure. By the end of year five, the estimated difference is still more than $2,000 per student but no longer statistically significant.[32]

Panels B and C of the table show results for the low and high quality samples respectively. Bond issues have larger effects on spending in districts with lower initial quality. In both samples, however, the pattern of spending mirrors the full sample results, with the gap in cumulative expenditures growing through year three, shrinking modestly in year four, and fading to statistical insignificance in year five.

These results guide my decision to look for effects on test scores four years after the bond election. It seems likely that any positive effect of construction on test scores would take some time to materialize; in fact, the initial impact of a major construction project could well be disruptive rather than beneficial. Four years is the longest lag at which there are large and strongly statistically significant differences in cumulative construction spending between districts where measures just barely passed and just barely failed. Moreover, in the fourth year after the bond election, average (annual) capital spending in districts that approved a bond issue has fallen most of the way back to its initial level, suggesting that whatever projects the bond issues funded are now complete. Cellini et al. (2010) find no effect of school construction on test scores, but Neilson and Zimmerman (2011), who do find that school construction raises test scores, find that the positive effects materialize as soon as the new buildings are finished.

Table 1.19 provides estimates of the effect of passing a bond measure on instructional spending. Legally, Texas school districts may use school bonds only to finance capital expenditures, but, since money is fungible, one might expect that districts would still use some of the bond proceeds for other purposes. The table shows that this does not seem to be the case.[33] The estimated impacts on instructional spending are small, mostly negative, and uniformly insignificant. Thus, the reduced form and two-stage least squares results should be interpreted as capturing exclusively the effects of additional construction spending.

[32] Because my data series ends in 2011. I observe different districts for different numbers of years after their school bond elections, and so Figure 1.14 and Table 1.18 show results for an unbalanced panel. The basic pattern of results is the same if I restrict the sample to districts that I observe for all six years.

[33] This result is consistent with the large literature on the "fly paper" effect, which finds that states and localities generally spend dedicated revenues and intergovernmental grants on their designated purpose. For an overview of this literature, see Hines and Thaler (1995).

1.7.3 Reduced Form and Two-Stage Least Squares Results

Figure 1.15 shows the effect of passing a bond measure on test score pass rates four years after the election for the full sample and for the low and high initial quality samples. In none of the three panels does there seem to be any discontinuity in test pass rates at the threshold for bond measure passage.

Table 1.20 provides the two-stage least squares results, parameterized to give the effect of a $1,000 per student increase in capital spending on average pass rates. Column 1 shows that there is no statistically significant effect on test pass rates in the full sample. The results are actually quite precise, largely because I control for test pass rates in the year before the election. I can rule out (at the 95% confidence level) effects as small as a 0.23 percentage point (0.04 standard deviation) increase in pass rates from a $1,000 increase in per-student spending.

Unsurprisingly given the lack of any treatment effect in the full sample, the results provide no basis for comparing the marginal product of capital spending in low- versus high-quality districts. The point estimates for both sub-samples are small and statistically insignificant. Column 4 shows that the coefficient of interest, the interaction between the treatment effect and having high initial quality is positive; the point estimate implies that a $1,000 increase in per-student spending has a 0.4 percentage point larger impact on test pass rates in a high- versus a low-initial quality district. However, this estimate is both small (roughly 0.08 standard deviations) and statistically insignificant.[34]

1.7.4 Discussion

It is possible that my findings indicate that the two thirds of Texas school districts that passed bond measures at some point during the past fifteen years were all making a mistake. It seems more likely, though, that I am simply testing for productivity effects using the wrong outcome measure.

I use standardized test performance as my outcome measure because it is an obvious objective measure of school productivity.[35] But improvements in test scores may not be the main benefit of increased capital spending. While Neilson and Zimmerman (2011) find that school construction in New Haven raises test scores, Cellini et al. (2010) find that school bond votes in California have no effect on test scores but sizable effects on home values. This suggests that homebuyers value school construction for some reason other than an impact on test performance, perhaps because it makes their children's school experience more enjoyable or because it enhances the aesthetics of the neighborhood.

[34] I also tried looking at the effect of capital spending on dropout rates. Again, neither the full sample treatment effect estimate nor the estimated interaction between the treatment effect and initial quality is statistically significant. Unsurprisingly, these results were also much less precise than the test score results, since there is much less variation in dropout rates, especially in the high initial quality sample. In the full sample, I cannot rule out the possibility that a $1,000 increase in per-student spending reduces dropout rates by as much as 0.2 percentage points; the mean dropout rate in the full sample is about 1 percentage point.

[35] I use pass rates rather than mean scores because of data constraints. A better measure would be average test scores, since it would incorporate the performance of all students, rather than just those near the threshold.

One possible direction for future research, therefore, would be to implement the same test for the correlation between the marginal productivity of funds and initial quality but use housing values in place of test scores as the productivity measure. This change would improve the test by using a more comprehensive measure of the effects of new spending. In fact, theoretically, the effect on home values is the best possible measure of the marginal productivity of additional capital spending since, under standard rationality assumptions, changes in home values should reflect all benefits of the new construction (test scores gains, amenity value, etc.).

It may seem odd to test for the efficiency of voting outcomes based on the assumption that housing values are efficiently determined. All that is required for this approach to make sense, however, is that individuals be substantially more likely to make socially efficient decisions when buying a home than when voting on a ballot measure. This seems plausible for two reasons. First, individuals likely devote considerably more attention and investigative effort to buying a house than to voting in a school bond election. Second, home values will be efficient as long as each individual correctly accounts for his own preferences, whereas efficient voting outcomes depend on appropriately aggregating everyone's preferences.

1.8 Conclusion

This paper investigates how voters make decisions about whether to increase local taxes and spending. Using variation generated by Texas's school district ratings system, I test the hypothesis that voters are more likely to support additional spending on schools when they perceive current service quality to be high. I find that support for school bond measures is 10 percentage points higher in districts that receive an exemplary versus a recognized rating. Voters respond to the level of the district's rating, not just to whether the most recent rating is better or worse than the previous year's.

From the standpoint of economic efficiency, decisions about new spending should be based on the marginal product of funds, not the total product (i.e. quality). In Section 1.7 above, I attempt to test whether the total product predicts the marginal product in this setting by estimating the correlation between initial quality and the effect of passing a bond measure on standardized test outcomes. I find that passing a bond measure has no effect on test scores in either the high or the low initial quality subsample, suggesting that test scores may be the wrong outcome for measuring the productivity of new capital spending. One direction for future work would be to improve this test for efficiency by substituting a more comprehensive measure of the effect of bond measures, perhaps the effect on housing values.

Another direction for future work would be to examine whether the school district ratings affect voters' preferences about funding for other levels of government. In particular, during my sample period, Texas voters had the opportunity to vote on several state constitutional amendments dealing with funding issues. Employing the same research design I have used here, one could examine whether perceptions of local school quality affect voters' support for increasing or reducing state taxes and spending. If local school ratings did affect votes in statewide elections, this would suggest that voters make funding decisions at least in part based on their general sense of how government – at all levels – is performing, rather than by reference to the specific policy under consideration.

References

Cabral, Marika and Caroline Hoxby, "The Hated Property Tax: Salience, Tax Rates, and Tax Revolts," 2011.

Campbell, Andrea, "What Americans Think of Taxes," in Isaac William Martin, Ajay K. Mehrotra, and Monica Prasad, eds., *The New Fiscal Sociology: Taxation in Comparative and Historical Perspective*, Cambridge University Press, 2009, pp. 48–67.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein, "The Value of School Facility Investments: Evidence From a Dynamic Regression Discontinuity Design," *Quarterly Journal of Economics*, 2010, *125* (1), 215–261.

Figlio, David N. and Cecelia Elena Rouse, "Do Accountability and Voucher Threats Improve Low-Performing Schools?," *Journal of Public Economics*, 2006, *90*, 239–255.

_ **and Maurice E. Lucas**, "What's in a Grade? School Report Cards and the Housing Market," *American Economic Review*, 2004, *94* (3), 591–604.

Finkelstein, Amy, "E-ZTax: Tax Salience and Tax Rates," *Quarterly Journal of Economics*, 2009, *124* (3), 969–1010.

Hines, James R. and Richard H. Thaler, "Anomalies: The Flypaper Effect," *Journal of Economic Perspectives*, 1995, *9* (4), 217–226.

Hoxby, Caroline, "All School Finance Equalizations Are Not Created Equal," *Quarterly Journal of Economics*, 2001, *116* (4), 1189–1231.

Imbens, Guido and Karthik Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *National Bureau of Economic Research Working Paper*, 2009.

Lacetera, Nicola, Devin G. Pope, and Justin R. Sydnor, "Heuristic Thinking and Limited Attention in the Car Market," *American Economic Review*, Forthcoming.

Luca, Michael, "Reviews, Reputations, and Revenue: The Case of Yelp," *Harvard Business School Working Paper*, 2011.

_ **and Jonathan Smith**, "Salience in Quality Disclosure: Evidence from the U.S. News College Rankings," *Harvard Business School Working Paper*, 2011.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142*, 698–714.

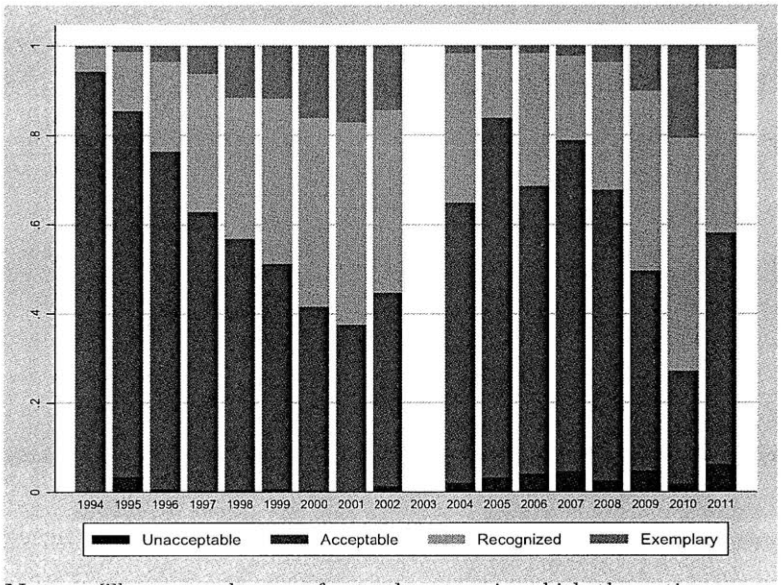
Neilson, Christopher and Seth Zimmerman, "The Effect of School Construction on Test Scores, School Enrollment, and Home Prices," *IZA Discussion Paper*, 2011.

Papay, John P., John B. Willet, and Richard J. Murnane, "Extending the Regression Discontinuity Approach to Multiple Assignment Variables," *Journal of Econometrics*, 2011, *161*, 203–207.

Pope, Devin G., "Reacting to Rankings: Evidence from America's Best Hospitals," *Journal of Health Economics*, 2009, *28*, 1154–1165.

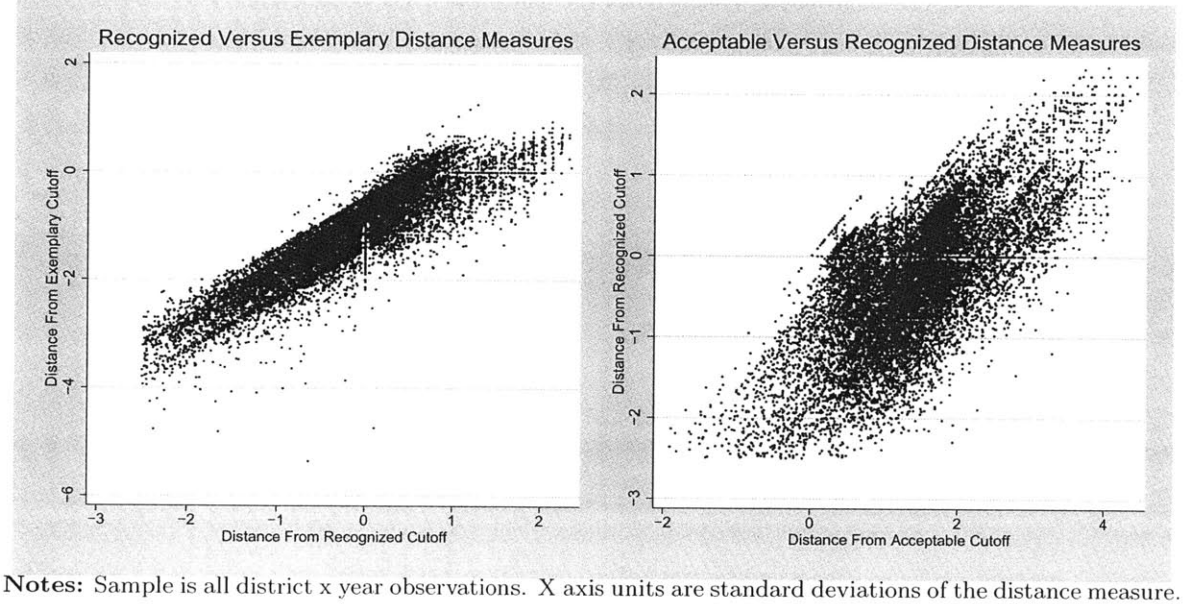
Reback, Randall, "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics*, 2008, *92*, 1394–1415.

Figure 1.1: Share of Districts Receiving Each Rating By Year



Notes: The years shown refer to the years in which the ratings were issued. No ratings were issued in 2003.

Figure 1.2: Non-Linearity of the Ratings System



Notes: Sample is all district x year observations. X axis units are standard deviations of the distance measure. Data points represent means for 0.1 standard deviation cells.

Figure 1.3: Texas Education Agency Ratings Formula Summary for 2010-2011

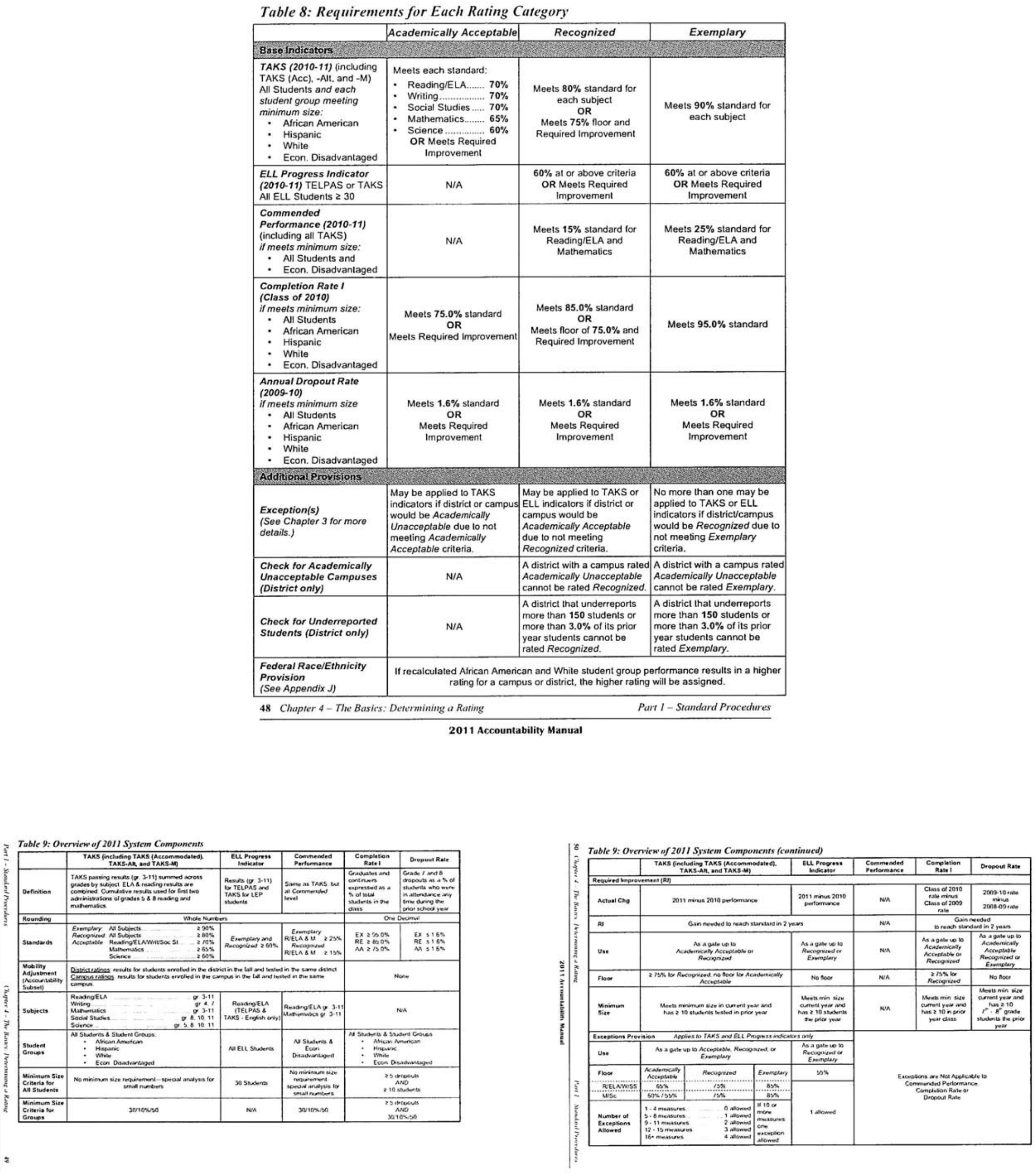


Table 8: Requirements for Each Rating Category

	Academically Acceptable	*Recognized*	*Exemplary*
Base Indicators			
TAKS (2010-11) (including TAKS (Acc), -Alt, and -M) All Students *and each student group meeting minimum size:* • African American • Hispanic • White • Econ. Disadvantaged	Meets each standard: • Reading/ELA....... **70%** • Writing................. **70%** • Social Studies..... **70%** • Mathematics........ **65%** • Science............... **60%** **OR** Meets Required Improvement	Meets **80%** standard for each subject **OR** Meets **75%** floor and Required Improvement	Meets **90%** standard for each subject
ELL Progress Indicator (2010-11) TELPAS or TAKS All ELL Students ≥ 30	N/A	**60%** at or above criteria **OR** Meets Required Improvement	**60%** at or above criteria **OR** Meets Required Improvement
Commended Performance (2010-11) (including all TAKS) *if meets minimum size:* • All Students and • Econ. Disadvantaged	N/A	Meets **15%** standard for Reading/ELA and Mathematics	Meets **25%** standard for Reading/ELA and Mathematics
Completion Rate I (Class of 2010) *if meets minimum size:* • All Students • African American • Hispanic • White • Econ. Disadvantaged	Meets **75.0%** standard **OR** Meets Required Improvement	Meets **85.0%** standard **OR** Meets floor of **75.0%** and Required Improvement	Meets **95.0%** standard
Annual Dropout Rate (2009-10) *if meets minimum size* • All Students • African American • Hispanic • White • Econ. Disadvantaged	Meets **1.6%** standard **OR** Meets Required Improvement	Meets **1.6%** standard **OR** Meets Required Improvement	Meets **1.6%** standard **OR** Meets Required Improvement
Additional Provisions			
Exception(s) *(See Chapter 3 for more details.)*	May be applied to TAKS indicators if district or campus would be *Academically Unacceptable* due to not meeting *Academically Acceptable* criteria.	May be applied to TAKS or ELL indicators if district or campus would be *Academically Acceptable* due to not meeting *Recognized* criteria.	No more than one may be applied to TAKS or ELL indicators if district/campus would be *Recognized* due to not meeting *Exemplary* criteria.
Check for Academically Unacceptable Campuses (District only)	N/A	A district with a campus rated *Academically Unacceptable* cannot be rated *Recognized*.	A district with a campus rated *Academically Unacceptable* cannot be rated *Exemplary*.
Check for Underreported Students (District only)	N/A	A district that underreports more than **150** students or more than **3.0%** of its prior year students cannot be rated *Recognized*.	A district that underreports more than **150** students or more than **3.0%** of its prior year students cannot be rated *Exemplary*.
Federal Race/Ethnicity Provision *(See Appendix J)*	If recalculated African American and White student group performance results in a higher rating for a campus or district, the higher rating will be assigned.		

48 *Chapter 4 – The Basics: Determining a Rating* *Part 1 – Standard Procedures*

2011 Accountability Manual

Table 9: Overview of 2011 System Components

	TAKS (including TAKS (Accommodated), TAKS-Alt, and TAKS-M)	**ELL Progress Indicator**	**Commended Performance**	**Completion Rate I**	**Dropout Rate**
Definition	TAKS passing results (gr. 3-11) summed across grades by subject. ELA & reading results are combined. Cumulative results used for first two administrations of grades 5 & 8 reading and mathematics.	Results (gr. 3-11) for TELPAS and TAKS for LEP students	Same as TAKS, but at *Commended* level	Graduates and continuers expressed as a % of total students in the class	Grade 7 and 8 dropouts as a % of students who were in attendance any time during the prior school year
Rounding	Whole Numbers			One Decimal	
Standards	*Exemplary:* All Subjects .... ≥ 90% *Recognized:* All Subjects .... ≥ 80% *Acceptable:* Reading/ELA/Writ/Soc St .... ≥ 70% Mathematics .... ≥ 65% Science .... ≥ 60%	*Exemplary* and *Recognized* ≥ 60%	*Exemplary* R/ELA & M ≥ 25% *Recognized* R/ELA & M ≥ 15%	EX ≥ 95.0% RE ≥ 85.0% AA ≥ 75.0%	EX ≤ 1.6% RE ≤ 1.6% AA ≤ 1.6%
Mobility Adjustment (Accountability Subset)	District ratings: results for students enrolled in the district in the fall and tested in the same district Campus ratings: results for students enrolled in the campus in the fall and tested in the same campus.			None	
Subjects	Reading/ELA .... gr. 3-11 Writing .... gr. 4, 7 Mathematics .... gr. 3-11 Social Studies .... gr. 8, 10, 11 Science .... gr. 5, 8, 10, 11	Reading/ELA (TELPAS & TAKS - English only)	Reading/ELA gr. 3-11 Mathematics gr. 3-11	N/A	
Student Groups	All Students & Student Groups: • African American • Hispanic • White • Econ. Disadvantaged	All ELL Students	All Students & Econ. Disadvantaged	All Students & Student Groups • African American • Hispanic • White • Econ. Disadvantaged	
Minimum Size Criteria for All Students	No minimum size requirement – special analysis for small numbers	30 Students	No minimum size requirement special analysis for small numbers	≥ 5 dropouts *AND* ≥ 10 students	
Minimum Size Criteria for Groups	30/10%/50	N/A	30/10%/50	≥ 5 dropouts *AND* 30/10%/50	

Part 1 – Standard Procedures

2011 Accountability Manual

Chapter 4 – The Basics: Determining a Rating

49

50 *Chapter 4 – The Basics: Determining a Rating*

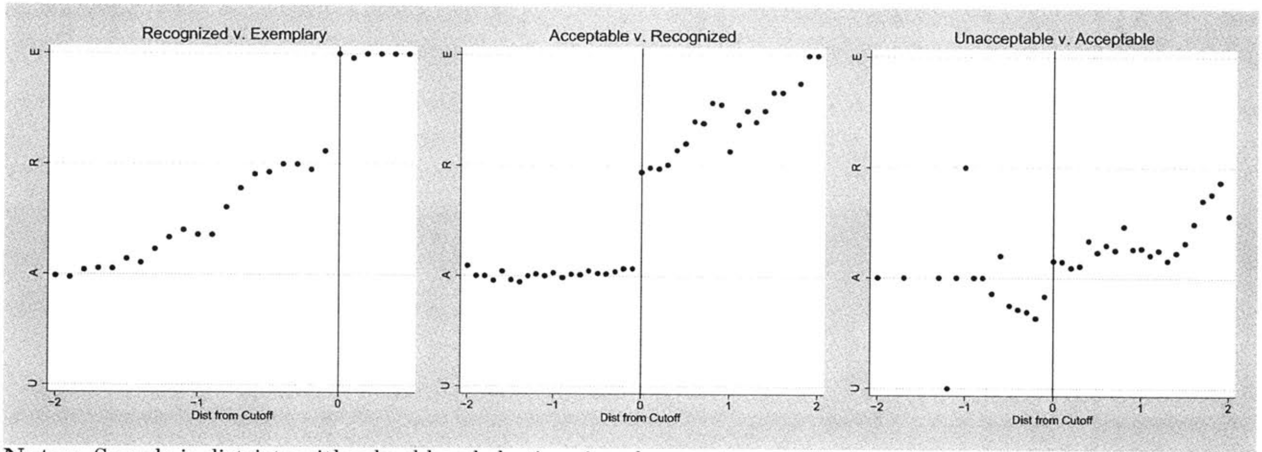
Table 9: Overview of 2011 System Components (continued)

	TAKS (including TAKS (Accommodated), TAKS-Alt, and TAKS-M)			**ELL Progress Indicator**	**Commended Performance**	**Completion Rate I**	**Dropout Rate**
Required Improvement (RI)							
Actual Chg	2011 minus 2010 performance			2011 minus 2010 performance	N/A	Class of 2010 rate minus Class of 2009 rate	2009-10 rate minus 2008-09 rate
RI	Gain needed to reach standard in 2 years				N/A	Gain needed to reach standard in 2 years	
Use	As a gate up to *Academically Acceptable* or *Recognized*			As a gate up to *Recognized* or *Exemplary*	N/A	As a gate up to *Academically Acceptable* or *Recognized*	As a gate up to *Academically Acceptable Recognized* or *Exemplary*
Floor	≥ 75% for *Recognized*, no floor for *Academically Acceptable*			No floor	N/A	≥ 75% for *Recognized*	No floor
Minimum Size	Meets minimum size in current year and has ≥ 10 students tested in prior year			Meets min. size current year and has ≥ 10 students the prior year	N/A	Meets min. size current year and has ≥ 10 in prior year class	Meets min. size current year and has ≥ 10 7^{th} - 8^{th} grade students the prior year
Exceptions Provision	*Applies to TAKS and ELL Progress indicators only*				Exceptions are Not Applicable to Commended Performance, Completion Rate or Dropout Rate		
Use	As a gate up to *Acceptable*, *Recognized*, or *Exemplary*			As a gate up to *Recognized* or *Exemplary*			
Floor	*Academically Acceptable*	*Recognized*	*Exemplary*	55%			
R/ELA/W/SS	65%	75%	85%				
M/Sc	60% / 55%	75%	85%				
Number of Exceptions Allowed	1 - 4 measures .... 0 allowed 5 - 8 measures .... 1 allowed 9 - 11 measures .... 2 allowed 12 - 15 measures .... 3 allowed 16+ measures .... 4 allowed		If 10 or more measures, one exception allowed	1 allowed			

2011 Accountability Manual

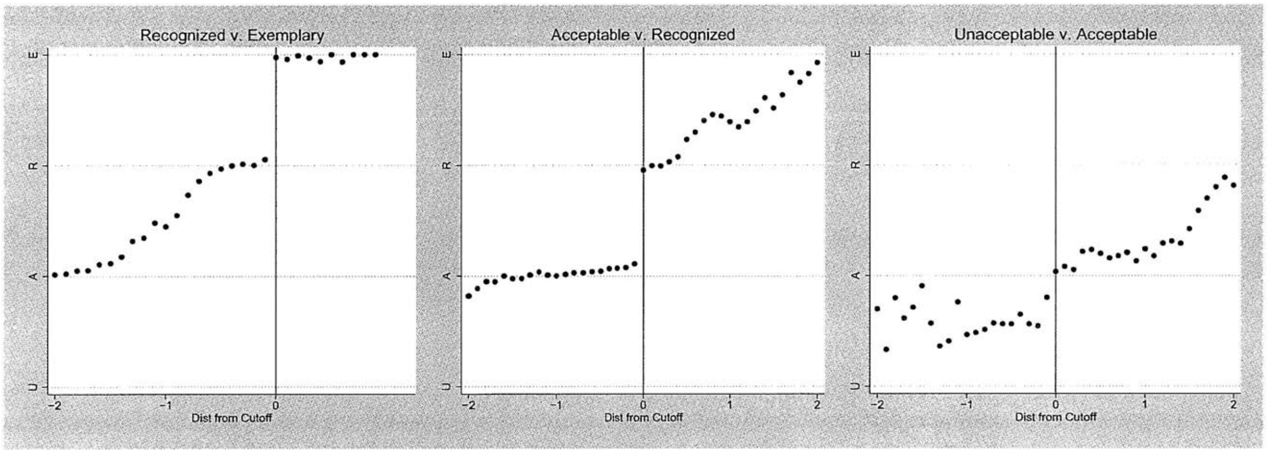
Part 1 – Standard Procedures

Figure 1.4: Distance Measures Versus Actual Ratings - Vote Sample



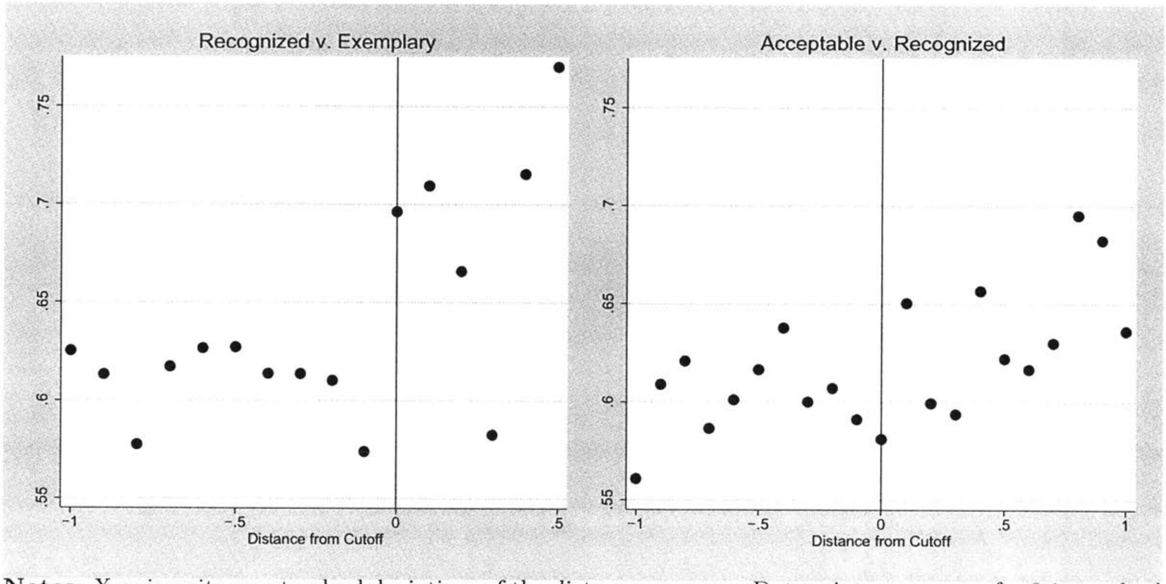
Notes: Sample is districts with school bond elections in subsequent year. X axis units are standard deviations of the distance measure. Data points represent means for 0.1 standard deviation cells. Ratings are converted to numbers as follows: unacceptable = 0, acceptable = 1, recognized = 2, exemplary = 3. 94% of constructed ratings match actual ratings (97% where constructed rating is not unacceptable).

Figure 1.5: Distance Measures Versus Actual Ratings - Full Sample



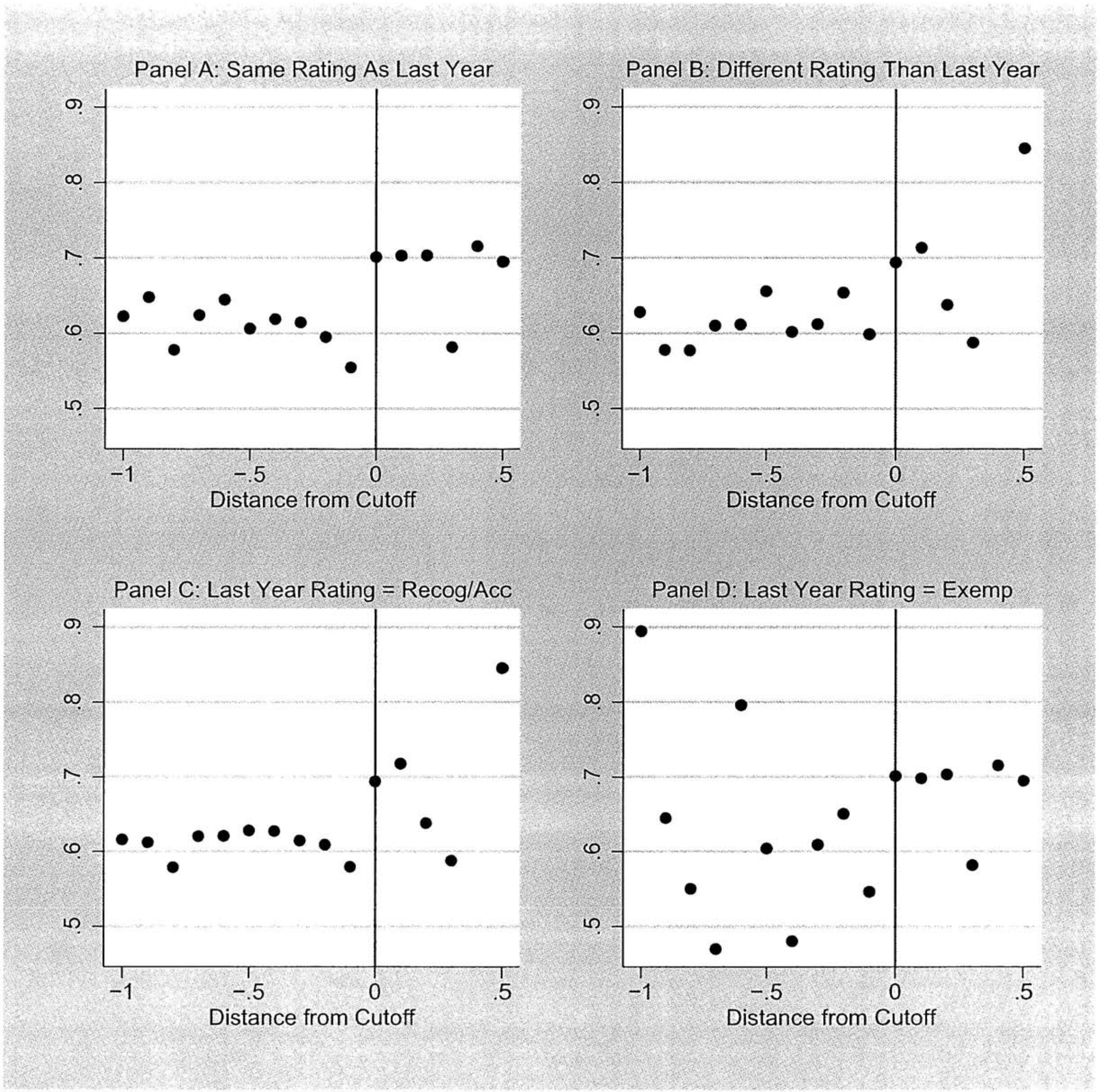
Notes: Sample is all district x year observations. X axis units are standard deviations of the distance measure. Data points represent means for 0.1 standard deviation cells. Ratings are converted to numbers as follows: unacceptable = 0, acceptable = 1, recognized = 2, exemplary = 3. 90% of constructed ratings match actual ratings (94% where constructed rating is not unacceptable).

Figure 1.6: Yes-Vote Share Around Ratings Cutoffs



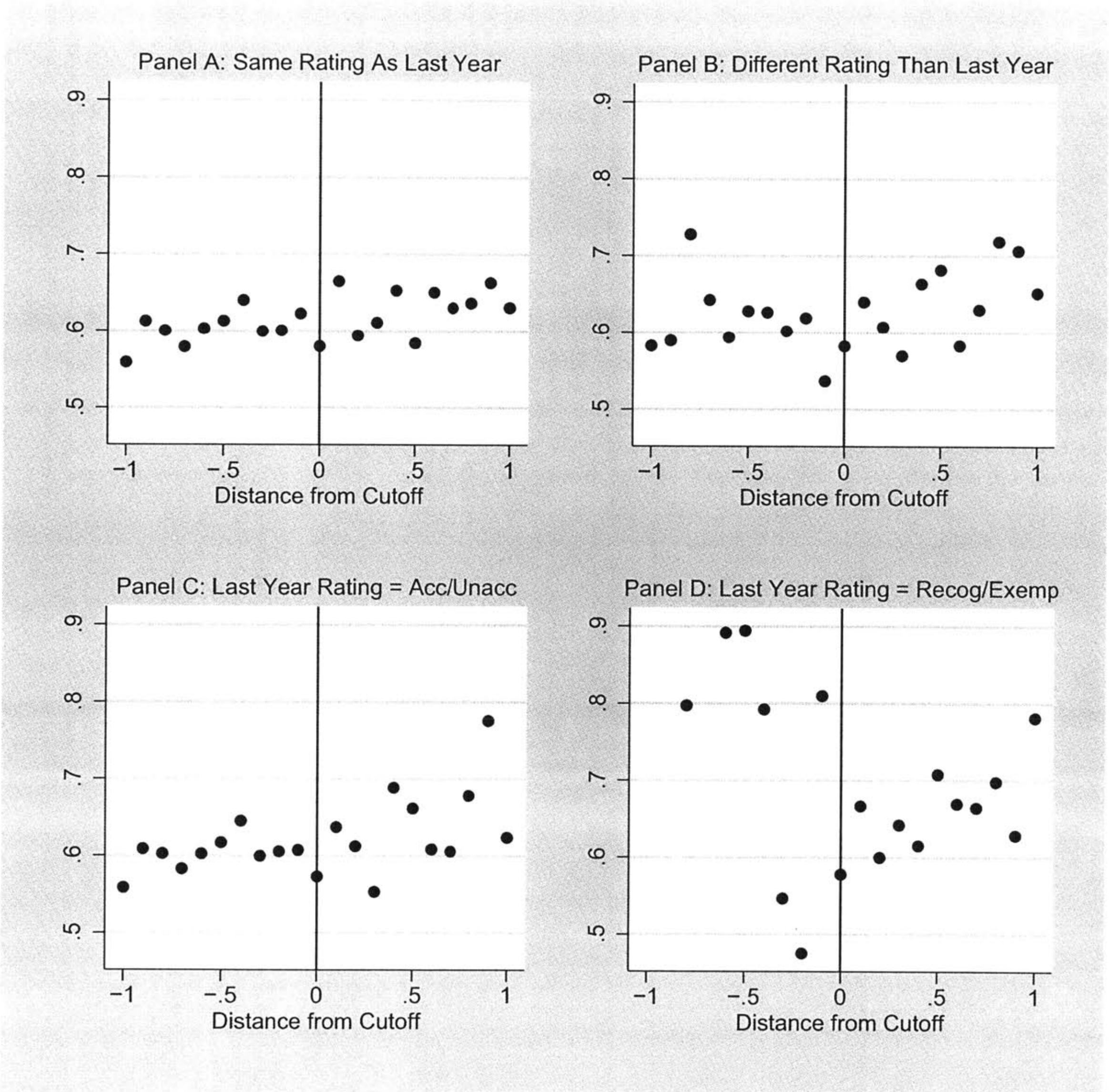
Notes: X-axis units are standard deviations of the distance measure. Datapoints are means for 0.1 standard deviation cells. Sample size in left panel is 887 districts; sample size in right panel is 1,693 districts.

Figure 1.7: Yes-Vote Share Around Exemplary Cutoff By Prior-Year Rating



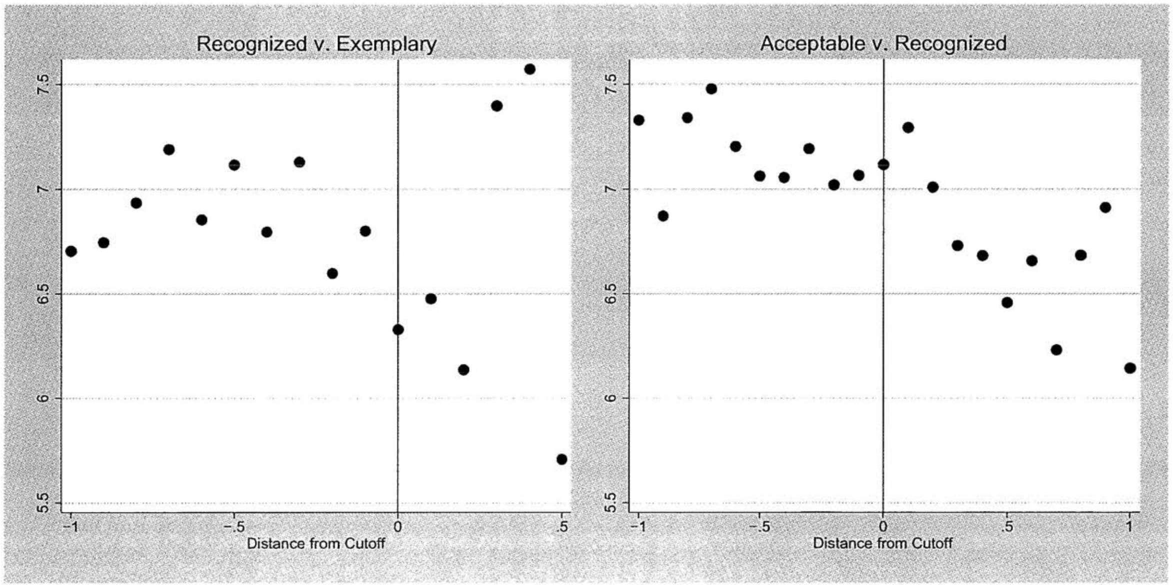
Notes: X-axis units are standard deviations of the distance measure. Datapoints are means for 0.1 standard deviation cells. Sample sizes: Panel A - 485 districts; Panel B - 402 districts; Panel C - 795 districts; Panel D - 92 districts.

Figure 1.8: Yes-Vote Share Around Recognized Cutoff By Prior-Year Rating



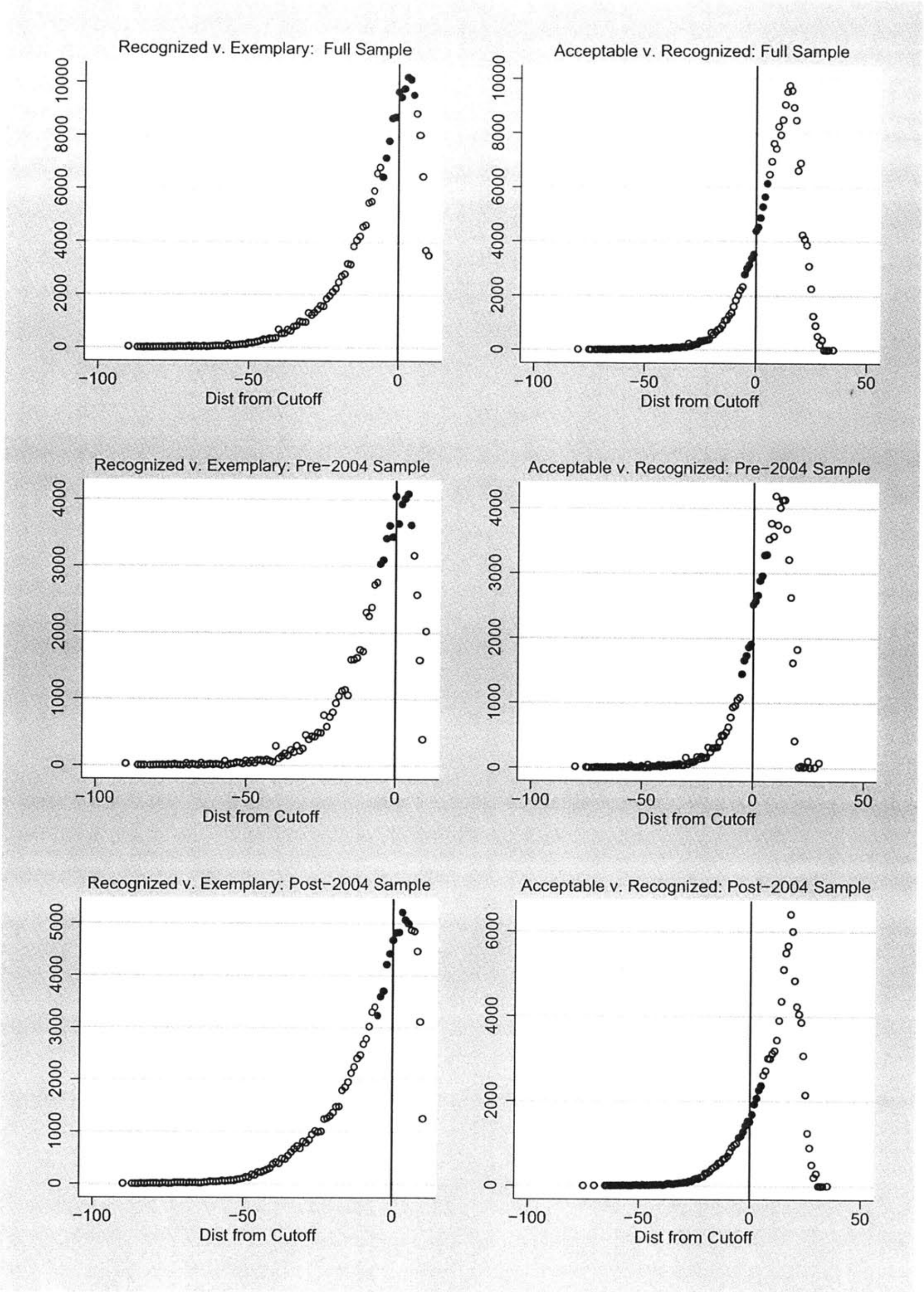
Notes: X-axis units are standard deviations of the distance measure. Datapoints are means for 0.1 standard deviation cells. Sample sizes: Panel A - 1,132 districts; Panel B - 561 districts; Panel C - 1,043 districts; Panel D - 650 districts.

Figure 1.9: Log Voter Turnout



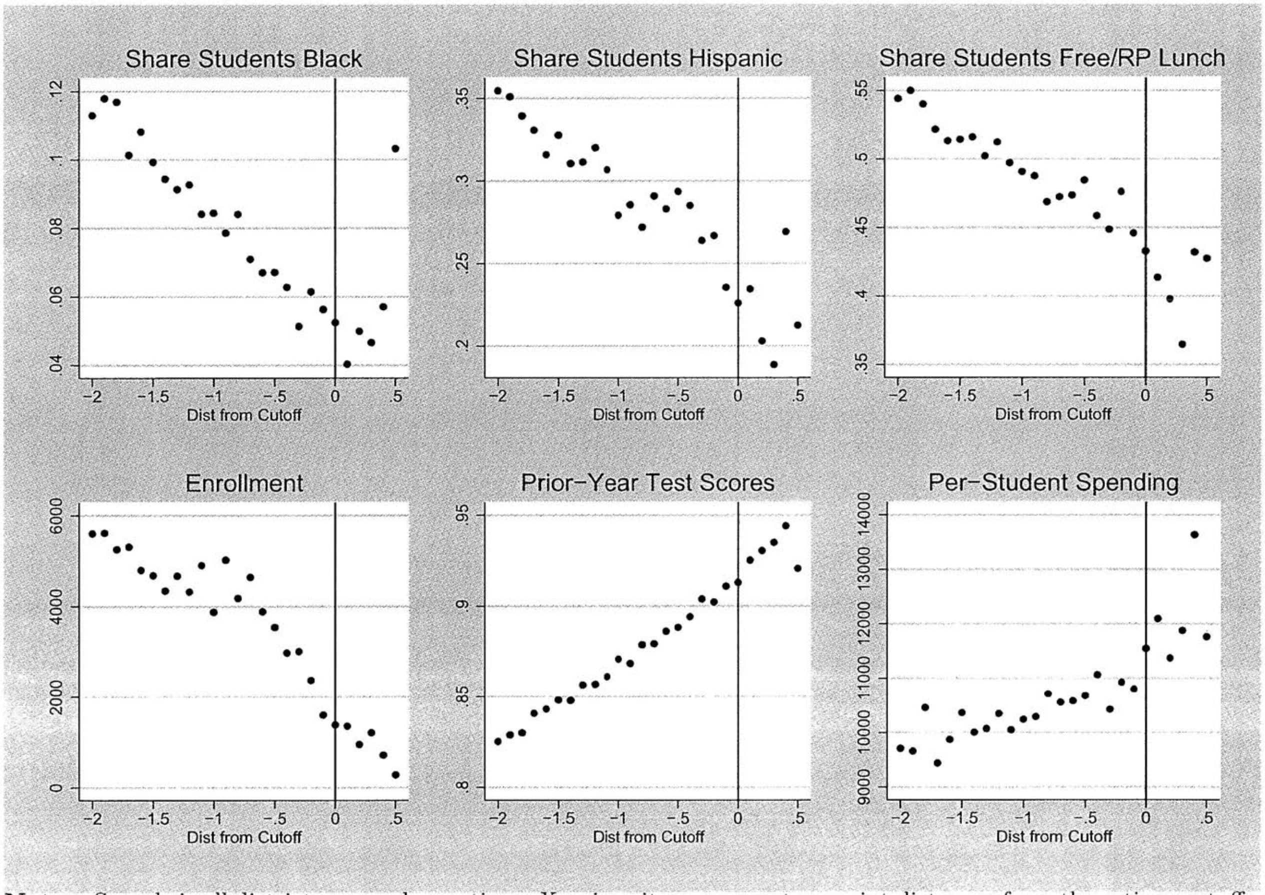
Notes: Log voter turnout is the log of the number of individuals voting in the election. X-axis units are standard deviations of the distance measure. Datapoints are means for 0.1 standard deviation cells. Sample size in left panel is 887 districts; sample size in right panel is 1,693 districts.

Figure 1.10: Number of Standardized Test Indicators By Distance From the Ratings Cutoffs



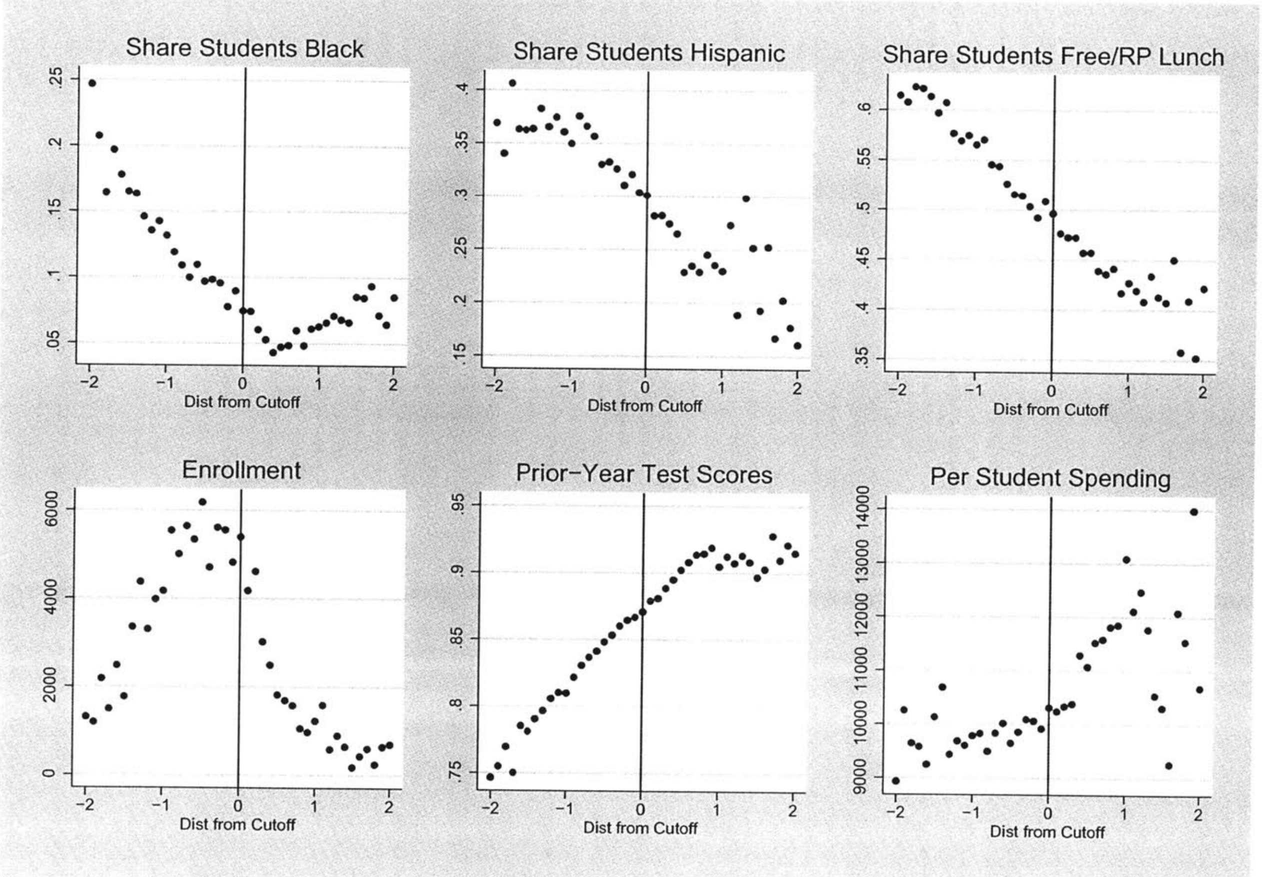
Notes: Sample is all standardized test pass rate indicators included in the ratings formula (i.e. standardized test pass rates for all students and for the four subgroups for those subjects included in the formula in a given year). X-axis units are percentage point distances from the ratings cutoffs. Y-axis units are numbers of indicators. Datapoints within 5 percentage points of the cutoffs are shown as solid red circles, while datapoints further from the cutoff are shown as hollow blue circles.

Figure 1.11: Covariate Balance Around the Exemplary Cutoff



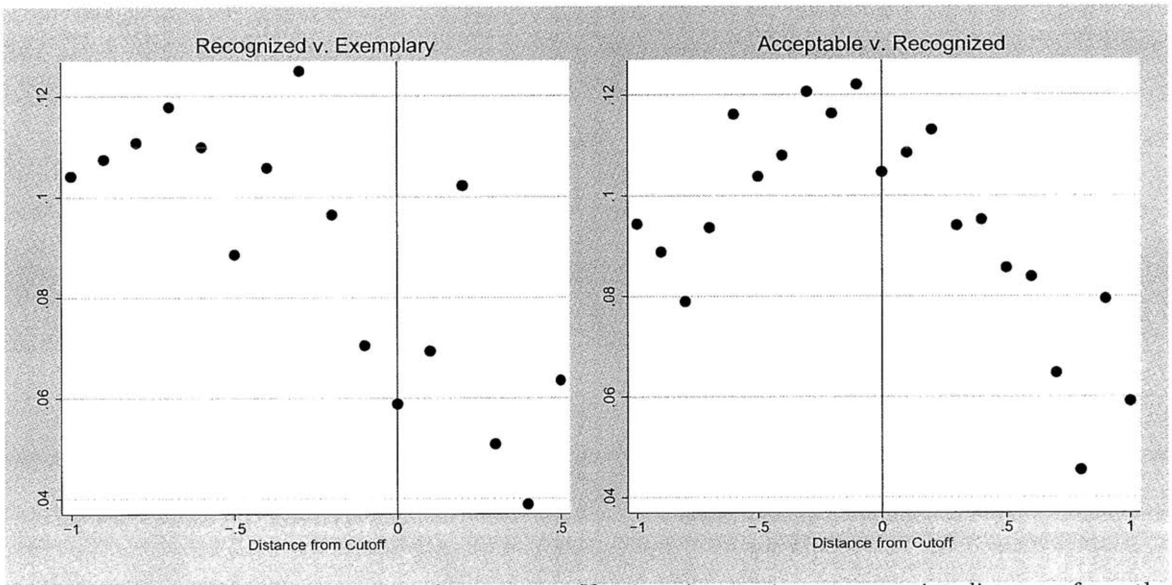
Notes: Sample is all district x year observations. X-axis units are percentage point distances from the ratings cutoffs. Datapoints are cell means for 0.1 standard deviation cells. Data are purged of year effects.

Figure 1.12: Covariate Balance Around the Recognized Cutoff



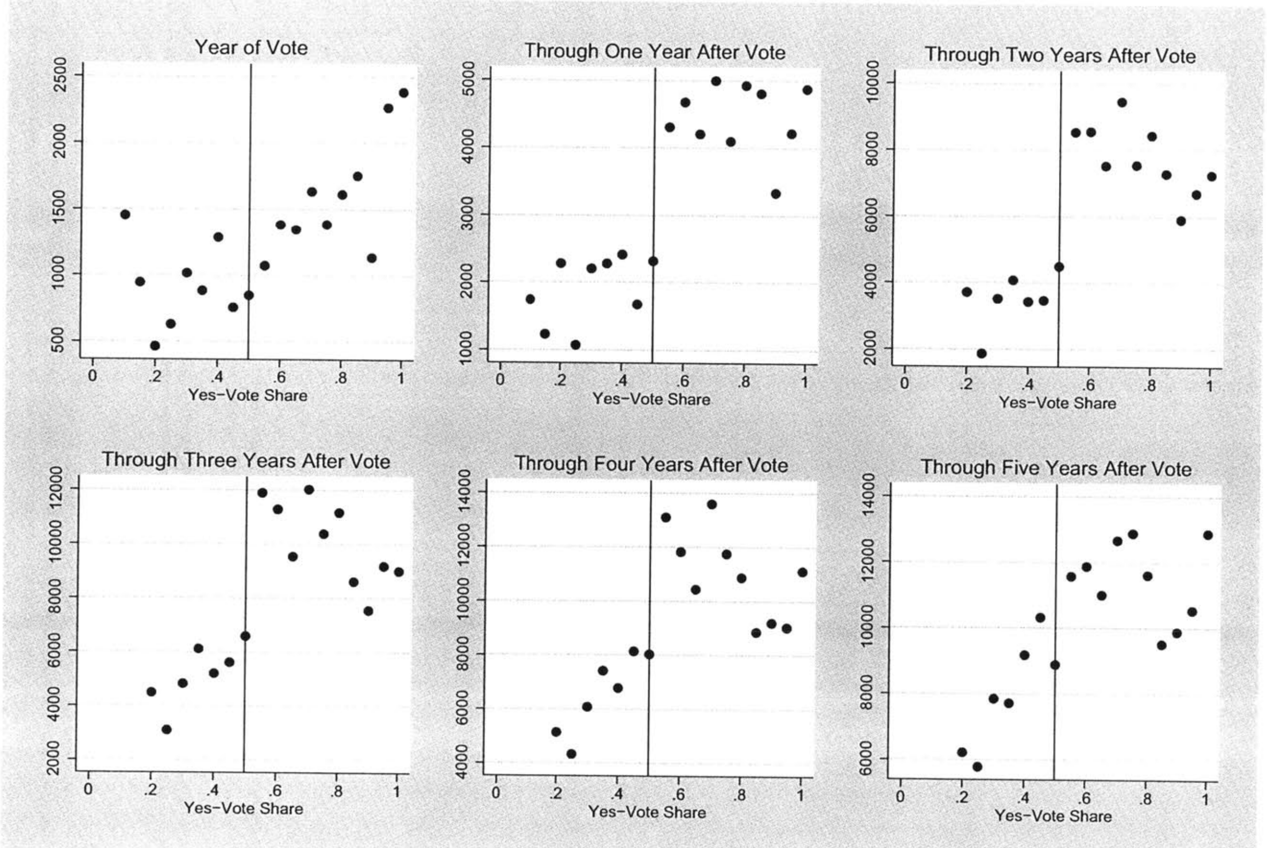
Notes: Sample is all district x year observations. X-axis units are percentage point distances from the ratings cutoffs. Datapoints are cell means for 0.1 standard deviation cells. Data are purged of year effects.

Figure 1.13: Share of Districts Holding Votes



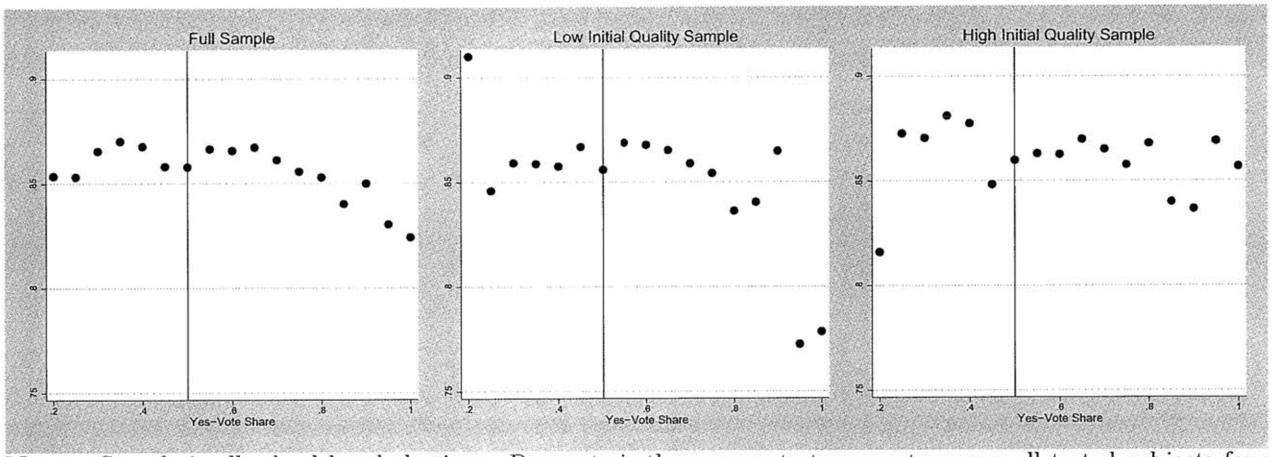
Notes: Sample is all district x year observations. X-axis units are percentage point distances from the ratings cutoffs. Datapoints are cell means for 0.1 standard deviation cells.

Figure 1.14: Capital Expenditures Per Student



Notes: Sample is all school bond elections. Capital expenditures are cumulative and include the year of the vote plus the specified number of subsequent years. X-axis units are percentage points of yes-vote share. Datapoints are cell means for 5 percentage point bins of yes-vote share. The vertical line is drawn at 50%, so means to the right of the line are for districts where bond measures passed.

Figure 1.15: Average Pass Rate on State Standardized Tests Four Years After School Bond Election



Notes: Sample is all school bond elections. Pass rate is the average test pass rate across all tested subjects four years after the election year. X-axis units are percentage points of yes-vote share. Datapoints are cell means for 5 percentage point bins of yes-vote share. The vertical line is drawn at 50%, so means to the right of the line are for districts where bond measures passed.

Table 1.1: Ratings Transition Matrix

	Rating in Year t+1			
Rating in Year t	Exemplary	Recognized	Acceptable	Unacceptable
Exemplary	55.3	36.3	8.2	0.0
Recognized	12.4	54.5	32.8	0.0
Acceptable	1.1	21.6	74.5	2.9
Unacceptable	1.0	10.7	68.6	19.7

Notes: The table shows the average shares of districts transitioning between ratings categories, omitting the 2002-2004 switch to the new ratings system. On average, 8% of districts are rated exemplary, 31% are rated recognized, 59% are rated acceptable, and 2% are rated unacceptable.

Table 1.2: District Summary Statistics

	Exemp	Recog	Acc	Unacc	All	W/Votes
Mean Enrollment	1,099	3,078	4,865	1,723	3,879	6,682
Med. Enrollment	352	751	1,122	392	842	1,804
Mean Share F/RP Lunch	0.40	0.47	0.55	0.72	0.52	0.47
Mean Share Black	0.05	0.06	0.11	0.27	0.09	0.09
Mean Share Hispanic	0.21	0.27	0.35	0.42	0.31	0.32
Mean Tests Pass Rate	0.96	0.91	0.83	0.68	0.86	0.87
Mean Spending/Student	14,031	12,616	11,411	11,663	12,040	11,542
Med. Spending/Student	11,246	10,614	10,184	10,514	10,413	9,905
N (districts x years)	1,387	5,507	9,207	388	17,198	1,577
N (districts)					1,303	782

Notes: Sample is all observations for 1996-2011. Districts with votes are those with a school bond vote at some point in the year following the reference ratings year. The state tests pass rate is computed as the average share of students in the district passing tests in reading, English language arts, math, and (when applicable) science and social studies.

Table 1.3: Election Summary Statistics

	Exemp	Recog	Acc	Unacc	All
Total Elections	116	692	1,245	17	2,070
Elections Per District Per Year	0.08	0.13	0.14	0.04	0.13
Elections Per District					1.6
Share of Bond Measures Passed	0.82	0.72	0.75	0.82	0.74
Mean Yes-Vote Share	0.67	0.61	0.60	0.62	0.61
Std. Yes-Vote Share	0.17	0.17	0.15	0.17	0.16
Mean Amt. Authorized (Millions of $2011)	22.0	44.2	50.8	54.0	47.0
Median Amt. Authorized (Millions of $2011)	6.3	1.2	1.5	1.6	1.3
Mean Turnout	1,597	2,364	2,300	3,400	2,892
Median Turnout	639	912	1,124	1,192	1,010
N (districts x years)	1,387	5,507	9,207	388	17,198
N (districts)					1,303

Notes: Sample is all observations for 1996-2011. Amount authorized is the bond issue amount authorized by the ballot measure. Yes-vote share is the share of voters voting in favor of the bond measure.

Table 1.4: First Stage Results

	Dependent Variable: Numerical Rating					
	Vote Sample			Full Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Exemplary v. Recognized Cutoff						
Dist >= 0	0.939***	0.943***	0.951***	0.926***	0.926***	0.926***
	(0.0520)	(0.0554)	(0.0574)	(0.0168)	(0.0166)	(0.0163)
F-Stat	326.6	289.3	273.9	3047.5	3123.9	3215.5
N	405	405	405	3384	3384	3384
Panel B: Recognized v. Acceptable Cutoff						
Dist >= 0	0.838***	0.740***	0.864***	0.834***	0.774***	0.842***
	(0.0397)	(0.0423)	(0.0326)	(0.0132)	(0.0736)	(0.0121)
F-Stat	444.4	306.3	702.6	3979.3	110.5	4812.1
N	1108	405	1108	7396	3384	7396
Panel C: Acceptable v. Unacceptable Cutoff						
Dist >= 0	0.330	0.335	0.249	0.277***	0.271***	0.276***
	(0.246)	(0.239)	(0.247)	(0.0639)	(0.0596)	(0.0567)
F-Stat	1.800	1.954	1.014	18.75	20.63	23.67
N	284	284	284	1826	1826	1826
Year dummies	N	Y	Y	N	Y	Y
District covariates	N	N	Y	N	N	Y

Notes: Local linear regressions with bandwidth equal to 0.5 standard deviations of the distance measure. Vote sample is districts with bond votes; full sample is all district x year observations. Numerical ratings conversion: 0 = unacceptable, 1 = acceptable, 2 = recognized, 3 = exemplary. Standard errors in vote sample regressions are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.5: Effect of a Higher Rating on Yes-Vote Share

	Dependent Variable: Yes-Vote Share					
	Exemplary v. Recognized			Recognized v. Acceptable		
	(1)	(2)	(3)	(4)	(5)	(6)
Higher rating	0.107*	0.115**	0.0991**	0.00742	0.0281	0.0374
	(0.0421)	(0.0393)	(0.0383)	(0.0285)	(0.0275)	(0.0270)
Dist below cutoff	-0.102	-0.0875	-0.0562	-0.0943	-0.139*	-0.158*
	(0.0890)	(0.0858)	(0.0863)	(0.0625)	(0.0628)	(0.0622)
Dist above cutoff	0.0274	-0.0589	-0.110	0.183*	0.214*	0.235**
	(0.168)	(0.157)	(0.154)	(0.0879)	(0.0849)	(0.0824)
Constant	0.585***	0.610***	0.583***	0.586***	0.559***	0.549***
	(0.0271)	(0.0350)	(0.0440)	(0.0186)	(0.0199)	(0.0199)
N	405	405	405	1108	1108	1108
Year dummies	N	Y	Y	N	Y	Y
District covariates	N	N	Y	N	N	Y

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.6: Robustness of Results to Alternative Bandwidths and Specifications

	Dependent Variable: Yes-Vote Share					
	Local Linear Regressions				Polynomial Regressions	
	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	BW = 1.0σ	BW = 0.7σ	BW = 0.3σ	Quadratic	Cubic
Panel A: Exemplary v. Recognized Cutoff						
Higher rating	0.0991**	0.0965*	0.0820*	0.129*	0.163**	0.179**
	(0.0383)	(0.0433)	(0.0360)	(0.0509)	(0.0523)	(0.0569)
Constant	0.583***	0.574***	0.585***	0.553***	0.528***	0.510***
	(0.0440)	(0.0365)	(0.0340)	(0.0549)	(0.0323)	(0.0407)
N	405	887	598	250	2070	2070
Panel B: Recognized v. Acceptable Cutoff						
Higher rating	0.0374	0.00592	0.0219	0.0264	0.00641	0.00896
	(0.0270)	(0.0213)	(0.0240)	(0.0340)	(0.0208)	(0.0247)
Constant	0.549***	0.592***	0.580***	0.557***	0.595***	0.573***
	(0.0199)	(0.0126)	(0.0157)	(0.0275)	(0.0133)	(0.0156)
N	1108	1693	1420	708	2070	2070

Notes: Two-stage least squares regressions. Bandwidths for local linear regressions are denominated in standard deviations of the distance measure. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Table 1.7: Effect of a Higher Rating on Bond Measure Pass Rates

	Dep Var: Bond Measure Passed	
	(1)	(2)
	Exemp v. Recog	Recog v. Acc
Higher rating	0.138	0.00701
	(0.0936)	(0.0726)
Dist below cutoff	0.0265	-0.350*
	(0.259)	(0.153)
Dist above cutoff	-0.617	0.546*
	(0.434)	(0.224)
Constant	0.708***	0.661***
	(0.113)	(0.0545)
N	405	1108

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. Dependent variable is a dummy for whether the bond measure passed. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.8: Effect of a Higher Rating on Yes-Vote Share By Prior-Year Rating

	Dependent Variable: Yes-Vote Share				
	(1)	(2)	(3)	(4)	(5)
	Baseline	Last Yr Same	Last Yr Dif	Improvers	Slippers
Panel A: Exemplary v. Recognized Cutoff					
Higher rating	0.0991**	0.0960*	0.0342	0.104*	0.109
	(0.0383)	(0.0456)	(0.0760)	(0.0496)	(0.142)
Dist below cutoff	-0.0562	-0.0733	0.0919	-0.0496	-0.154
	(0.0863)	(0.102)	(0.186)	(0.0897)	(0.425)
Dist above cutoff	-0.110	-0.0666	-0.282	-0.172	0.261
	(0.154)	(0.191)	(0.281)	(0.248)	(0.465)
Constant	0.583***	0.652***	0.701***	0.537***	0.522***
	(0.0440)	(0.0297)	(0.0556)	(0.0281)	(0.138)
N	405	248	157	331	74
Panel B: Recognized v. Acceptable Cutoff					
Higher rating	0.0374	0.0333	0.0752	0.0233	0.0454
	(0.0270)	(0.0364)	(0.0484)	(0.0349)	(0.0453)
Dist below cutoff	-0.158*	-0.0883	-0.280*	-0.171**	-0.164
	(0.0622)	(0.0623)	(0.134)	(0.0661)	(0.138)
Dist above cutoff	0.235**	0.170	0.338*	0.347**	0.190
	(0.0824)	(0.0975)	(0.155)	(0.124)	(0.138)
Constant	0.549***	0.566***	0.545***	0.556***	0.588***
	(0.0199)	(0.0214)	(0.0461)	(0.0228)	(0.0405)
N	1108	691	417	645	463

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. "Last yr same" sample is districts with the same rating as the year before. "Last yr dif" sample is districts with a different rating this year than last year. "Improvers" sample is districts that last year were below the relevant cutoff. "Slippers" sample is districts that last year were above the relevant cutoff. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.9: Effect of a Higher Rating on Yes-Vote Share By Time Since Ratings Were Issued

	Dependent Variable: Yes-Vote Share			
	(1) 0-3 Months	(2) 4-6 Months	(3) 7-9 Months	(4) 10-12 Months
Panel A: Exemplary v. Recognized Cutoff				
Higher rating	0.206*	0.184*	0.0392	0.0678
	(0.0824)	(0.0888)	(0.115)	(0.0708)
Dist below cutoff	-0.0460	-0.186	0.298	0.0238
	(0.241)	(0.119)	(0.213)	(0.164)
Dist above cutoff	-0.692	-0.198	-0.668	-0.145
	(0.448)	(0.304)	(0.353)	(0.214)
Constant	0.656***	0.512***	0.758***	0.650***
	(0.0835)	(0.0332)	(0.0977)	(0.0581)
N	101	84	68	152
Panel B: Recognized v. Acceptable Cutoff				
Higher rating	-0.0219	0.000696	0.0918	0.0388
	(0.0542)	(0.0503)	(0.0615)	(0.0425)
Dist below cutoff	-0.129	0.0123	-0.248	-0.181
	(0.121)	(0.0991)	(0.138)	(0.109)
Dist above cutoff	0.315	-0.00444	0.323	0.201
	(0.161)	(0.124)	(0.208)	(0.132)
Constant	0.635***	0.586***	0.629***	0.637***
	(0.0565)	(0.0287)	(0.0632)	(0.0292)
N	233	284	175	416

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. Sample is split by time between the release of the ratings and the bond election. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.10: Effect of a Higher Rating on Yes-Vote Share By Type of Election

	Dependent Variable: Yes-Vote Share		
	(1) High-Profile	(2) Statewide	(3) Other
Panel A: Exemplary v. Recognized Cutoff			
Higher rating	-0.0644	-0.0139	0.100*
	(0.0351)	(0.0470)	(0.0415)
Dist below cutoff	-0.317***	-0.0611	-0.0478
	(0.0558)	(0.144)	(0.0923)
Dist above cutoff	0.658***	-0.0122	-0.0472
	(0.0877)	(0.320)	(0.152)
Constant	0.526***	0.545***	0.615***
	(0.0229)	(0.0334)	(0.0416)
N	22	50	355
Panel B: Recognized v. Exemplary Cutoff			
Higher rating	-0.0592	-0.0115	0.0471
	(0.0455)	(0.0493)	(0.0305)
Dist below cutoff	0.0909	-0.000122	-0.207**
	(0.0886)	(0.109)	(0.0725)
Dist above cutoff	-0.0527	0.126	0.283**
	(0.127)	(0.167)	(0.0921)
Constant	0.646***	0.578***	0.627***
	(0.0420)	(0.0309)	(0.0231)
N	98	253	855

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. High profile elections are presidential, senatorial, and gubernatorial elections and presidential primaries. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.11: Effect of a Higher Rating on Log Voter Turnout

	Dep Var: Log Voter Turnout	
	(1)	(2)
	Exemp v. Recog	Recog v. Acc
Higher rating	-0.214	-0.113
	(0.239)	(0.149)
Dist below cutoff	-0.502	0.335
	(0.518)	(0.379)
Dist above cutoff	1.286	-0.713
	(0.940)	(0.499)
Constant	6.678***	7.441***
	(0.175)	(0.116)
N	405	1108

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. Dependent variable is the log of the number of individuals voting in the election. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.12: Testing for Bunching in Test Pass Rates

	Dependent Variable: Number of Indicators		
	(1)	(2)	(3)
	Full Sample	1996-2003	2004-2011
Panel A: Exemplary v. Recognized			
Dist >= 0	99.91	224.8	10.46
	(361.7)	(240.4)	(87.55)
Dist below cutoff	603.3***	133.1*	299.6***
	(78.95)	(42.94)	(13.12)
Dist above cutoff	-549.2**	-154.1	-221.9***
	(113.4)	(72.84)	(30.49)
Constant	9602.9***	3721.9***	4737.4***
	(322.5)	(177.0)	(61.69)
N	11	11	11
Panel B: Recognized v. Acceptable			
Dist >= 0	525.5**	381.5**	-81.86
	(109.9)	(78.40)	(35.49)
Dist below cutoff	184.7***	113.1***	102.7***
	(6.356)	(17.07)	(11.33)
Dist above cutoff	173.0**	36.04	72.27**
	(33.66)	(29.23)	(14.47)
Constant	3729.3***	2062.3***	1622.1***
	(22.33)	(49.16)	(27.29)
N	11	11	11

Notes: Local linear regressions with bandwidth equal to 5 percentage points. Sample is all standardized test pass rate indicators included in the ratings formula (i.e. standardized test pass rates for all students and for the four subgroups for those subjects included in the formula in a given year). Dependent variable is the number of pass rate indicators at a given distance from the cutoff. Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.13: Testing for Discontinuities in Covariates at the Exemplary Cutoff

	(1) Black	(2) Hispanic	(3) F/RP Lunch	(4) Enrollment	(5) Last-Yr Scores	(6) $/Student
Panel A: Full Sample						
Higher rating	-0.0111	-0.00571	-0.0119	-158.9	0.000212	674.0
	(0.00811)	(0.0177)	(0.0150)	(406.7)	(0.00281)	(628.3)
Constant	0.0573***	0.228***	0.433***	736.6*	0.931***	11874.7***
	(0.00726)	(0.0144)	(0.0119)	(351.2)	(0.00254)	(463.1)
N	3384	3384	3384	3384	3363	3377
Panel B: 1996-2003 Sample						
Higher rating	0.00269	-0.00924	-0.00293	-197.6	-0.00321	356.3
	(0.00649)	(0.0213)	(0.0174)	(372.5)	(0.00319)	(701.5)
Constant	0.0430***	0.228***	0.479***	222.2	0.933***	10998.3***
	(0.00669)	(0.0225)	(0.0176)	(419.4)	(0.00410)	(461.7)
N	2190	2190	2190	2190	2189	2190
Panel C: 2004-2011 Sample						
Higher rating	-0.0383	-0.00344	-0.0203	-395.3	0.00799	832.7
	(0.0202)	(0.0319)	(0.0281)	(947.9)	(0.00554)	(1322.8)
Constant	0.0888***	0.229***	0.392***	688.6	0.937***	13046.4***
	(0.0198)	(0.0281)	(0.0263)	(741.2)	(0.00549)	(1179.9)
N	1194	1194	1194	1194	1174	1187

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the running variable. Black, Hispanic, and free/RP lunch refer to the share of a district's students falling into these categories. Regressions are parameterized so that the constant coefficient gives the predicted value of the dependent variable just below the cutoff. All regressions include year effects. Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.14: Testing for Discontinuities in Covariates at the Recognized Cutoff

	(1)	(2)	(3)	(4)	(5)	(6)
	Black	Hispanic	F/RP Lunch	Enrollment	Last-Yr Scores	$/Student
Panel A: Full Sample						
Higher rating	-0.00434	-0.00369	-0.000788	530.2	-0.0000217	-46.50
	(0.00691)	(0.0145)	(0.0103)	(739.6)	(0.00285)	(323.5)
Constant	0.0815***	0.299***	0.490***	4940.6***	0.872***	10348.4***
	(0.00547)	(0.0111)	(0.00793)	(570.6)	(0.00216)	(214.8)
N	7396	7396	7396	7396	7380	7379
Panel B: 1996-2003 Sample						
Higher rating	-0.0124	-0.0129	-0.00807	874.0	0.00611	13.74
	(0.00857)	(0.0209)	(0.0145)	(1019.9)	(0.00396)	(323.8)
Constant	0.0905***	0.324***	0.492***	4593.1***	0.876***	10131.1***
	(0.00775)	(0.0195)	(0.0139)	(935.0)	(0.00396)	(252.1)
N	3879	3879	3879	3879	3878	3876
Panel C: 2004-2011 Sample						
Higher rating	0.00556	0.00609	0.00767	265.0	-0.00751	-96.39
	(0.0110)	(0.0201)	(0.0148)	(1080.6)	(0.00409)	(570.9)
Constant	0.0750***	0.274***	0.476***	5003.8***	0.874***	10867.1***
	(0.00981)	(0.0176)	(0.0129)	(859.7)	(0.00341)	(459.2)
N	3517	3517	3517	3517	3502	3503

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the running variable. Black, Hispanic, and free/RP lunch refer to the share of a district's students falling into these categories. Regressions are parameterized so that the constant coefficient gives the predicted value of the dependent variable just below the cutoff. All regressions include year effects. Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.15: Additional Robustness Checks

	Dependent Variable: Yes-Vote Share		
	(1) Baseline	(2) Post-2004 Sample	(3) Test Scores Only
Panel A: Exemplary v. Recognized			
Higher rating	0.0991**	0.292**	0.0887*
	(0.0383)	(0.0992)	(0.0429)
Dist below cutoff	-0.0562	-0.389*	-0.0196
	(0.0863)	(0.188)	(0.0585)
Dist above cutoff	-0.110	-0.116	-0.101
	(0.154)	(0.406)	(0.0866)
Constant	0.583***	0.500***	0.646***
	(0.0440)	(0.0824)	(0.0398)
N	405	79	279
Panel B: Recognized v. Acceptable			
Higher rating	0.0374	0.00227	0.0520*
	(0.0270)	(0.0352)	(0.0238)
Dist below cutoff	-0.158*	-0.00668	-0.127**
	(0.0622)	(0.0788)	(0.0409)
Dist above cutoff	0.235**	0.0557	0.160***
	(0.0824)	(0.133)	(0.0432)
Constant	0.549***	0.564***	0.538***
	(0.0199)	(0.0299)	(0.0198)
N	1108	425	1108

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the running variable. In the column labeled, "Test Scores Only," the distance measures are computed based only on test pass rates. Regressions are parameterized such that the constant coefficient gives the predicted value of the dependent variable just below the cutoff. All regressions include year effects and control for district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.16: Effect of a Higher Rating on Share of Districts Holding Votes

	Dep Var: Held Vote	
	(1) Exemp v. Recog	(2) Recog v. Acc
Higher rating	-0.0116	-0.00386
	(0.0190)	(0.0170)
Dist below cutoff	-0.0322	-0.0120
	(0.0445)	(0.0396)
Dist above cutoff	0.00118	-0.0174
	(0.0632)	(0.0495)
Constant	0.0597***	0.105***
	(0.0146)	(0.0127)
N	3384	7396

Notes: Two-stage least squares regressions with bandwidth equal to 0.5 standard deviations of the distance measure. Dependent variable is a dummy for whether the district held an election. Regressions are parameterized such that the constant coefficient gives the predicted value of yes-vote share just below the cutoff. All regressions include year effects and control for pre-election district characteristics. Robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.17: Summary Statistics for Districts with Close Elections Versus Other Districts with Votes

	Close Elections	Others W/Votes	P-Value: Different
Mean Enrollment	6,852	6,640	0.842
Med. Enrollment	2,209	1,717	0.027
Mean Share F/RP Lunch	0.48	0.47	0.812
Mean Share Black	0.09	0.08	0.210
Mean Share Hispanic	0.31	0.32	0.357
Mean Tests Pass Rate	0.86	0.87	0.357
Mean Spending/Student	10,823	11,721	0.002
Med. Spending/Student	9,785	9,935	0.288
Mean Capital Spending/Student	1,075	1,231	0.256
Med. Capital Spending/Student	478	513	0.512
N (districts x years)	315	1,262	1577
N (districts)	252	742	782

Notes: Close elections are those where the bond measure passed or failed by 5 percentage points or less. The state tests pass rate is computed as the average share of students in the district passing tests in reading, English language arts, math, and (when applicable) science and social studies.

Table 1.18: Effect of Passing a Bond Measure on Capital Spending

	Dep Var: Total Capital Expenditures Per Student Since Year of Vote					
	(1) Vote Year	(2) Year 1	(3) Year 2	(4) Year 3	(5) Year 4	(6) Year 5
Panel A: Full Sample						
Passed Vote	44.08	1419.3**	3097.0***	3970.9***	3351.1**	2267.3
	(175.4)	(453.6)	(797.0)	(1095.8)	(1255.4)	(1162.5)
Constant	859.6***	2309.6***	4465.9***	5789.4***	6362.2***	6862.2***
	(126.9)	(313.9)	(520.0)	(974.0)	(959.9)	(1128.1)
N	773	712	668	618	534	445
F-Stat	0.0632	9.792	15.10	13.13	7.125	3.804
Panel B: Low Initial Quality Sample						
Passed Vote	76.88	1778.1**	3749.1**	5213.4**	4479.3*	3368.6
	(215.0)	(655.1)	(1252.4)	(1885.7)	(2234.1)	(1929.2)
Constant	859.7***	2276.6***	4451.4***	6090.7***	8492.9***	7521.5**
	(171.6)	(475.3)	(1022.1)	(1541.5)	(1872.1)	(2273.0)
N	469	444	414	373	297	212
F-Stat	0.128	7.367	8.961	7.643	4.020	3.049
Panel C: High Initial Quality Sample						
Passed Vote	-62.55	891.2	2236.2*	2508.9*	2493.9*	2273.9
	(267.7)	(558.4)	(923.5)	(1148.2)	(1178.5)	(1383.5)
Constant	949.4***	3636.9***	4938.8***	7059.5***	8282.2***	9556.7**
	(230.5)	(921.1)	(1120.4)	(1843.2)	(2429.0)	(3192.4)
N	304	268	254	245	237	233
F-Stat	0.0546	2.547	5.863	4.774	4.478	2.702

Notes: Local linear regressions with bandwidth equal to 10 percentage points of yes-vote share. Dependent variable is cumulative per-student capital expenditures in the year of vote plus the specified number of subsequent years. The low initial quality sample consists of districts at least one standard deviation below the exemplary rating cutoff in the year before the bond vote. Regressions are parameterized such that the constant gives the predicted value of the outcome when yes-vote share is 50%. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.19: Effect of Passing a Bond Measure on Instructional Spending

	Dependent Variable: Instructional Expenditures Per Student					
	(1)	(2)	(3)	(4)	(5)	(6)
	Vote Year	Year 1	Year 2	Year 3	Year 4	Year 5
Panel A: Full Sample						
Passed Vote	-66.08	-76.88	-32.90	-10.10	-4.563	-86.79
	(41.04)	(58.57)	(71.86)	(59.80)	(73.85)	(82.27)
Constant	4652.3***	4725.1***	4747.5***	4846.6***	4838.4***	5050.0***
	(37.10)	(55.62)	(65.02)	(107.1)	(103.7)	(108.1)
N	773	712	668	618	534	445
Panel B: Low Initial Quality Sample						
Passed Vote	-92.97	-161.5	-2.803	-2.858	73.74	-159.5
	(48.28)	(82.76)	(101.7)	(86.09)	(101.9)	(137.9)
Constant	4810.6***	4969.7***	4928.5***	4876.5***	4939.9***	5023.0***
	(48.03)	(89.08)	(109.1)	(94.42)	(102.3)	(147.5)
N	469	444	414	373	297	212
Panel C: High Initial Quality Sample						
Passed Vote	-60.03	39.98	-59.18	-21.90	-39.22	-5.237
	(62.86)	(69.21)	(77.94)	(84.07)	(95.85)	(102.8)
Constant	4604.2***	4629.6***	4669.2***	4844.0***	4812.6***	4556.7***
	(81.65)	(88.47)	(127.0)	(166.2)	(143.1)	(209.0)
N	304	268	254	245	237	233

Notes: Local linear regressions with bandwidth equal to 10 percentage points of yes-vote share. Dependent variable is *annual* per-student instructional expenditures in specified year. The low initial quality sample consists of districts at least one standard deviation below the exemplary rating cutoff in the year before the bond vote. Regressions are parameterized such that the constant gives the predicted value of the outcome when yes-vote share is 50%. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.20: Effect of Passing a Bond Measure on Standardized Test Pass Rates

	Dep Var: Average Test Pass Rate After Four Years			
	(1)	(2)	(3)	(4)
	Full Sample	Low Quality	High Quality	Full Sample
Capital spending ($1,000s)	-0.00172	-0.00238	0.00168	-0.00238
	(0.00199)	(0.00203)	(0.00408)	(0.00203)
Yes-vote share	0.0973	0.144	0.0500	0.144
	(0.0803)	(0.154)	(0.0868)	(0.154)
Yes-vote share*passed vote	-0.0534	-0.105	-0.0730	-0.105
	(0.108)	(0.192)	(0.175)	(0.192)
Capital spending*high quality				0.00406
				(0.00455)
Yes-vote share*high quality				-0.0945
				(0.177)
Yes-vote share*passed vote*high quality				0.0321
				(0.260)
High quality				-0.0449
				(0.0477)
Constant	0.897***	0.841***	0.882***	0.841***
	(0.0174)	(0.0241)	(0.0442)	(0.0241)
N	497	279	218	497

Notes: Two-stage least squares regressions with bandwidth equal to 10 percentage points of yes-vote share and bond measure passage instrumenting for capital spending. Dependent variable is average (across subjects) pass rate on state standardized tests. Capital spending is a cumulative measure equal to per-student capital expenditures in the year of the bond measure plus the four subsequent years. The low initial quality sample consists of districts at least one standard deviation below the exemplary rating cutoff in the year before the bond vote. Regressions are parameterized such that the constant gives the predicted value of the outcome when yes-vote share is 50%. All regressions include year effects and control for pre-election district characteristics. Standard errors are clustered at the district x year level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 2

Moral Hazard in Health Insurance: How Important Is Forward Looking Behavior?

(Joint Work with Liran Einav, Amy Finkelstein, and Mark Cullen[1])

2.1 Introduction

The size and rapid growth of the healthcare sector – and the pressure this places on public sector budgets – has created great interest among both academics and policymakers in possible approaches to reducing healthcare spending. On the demand side, the standard, long-standing approach to constraining healthcare spending is through consumer cost sharing in health insurance, such as deductibles and coinsurance. Not surprisingly therefore, there is a substantial academic literature devoted to trying to quantify how the design of health insurance contracts affects medical spending. These estimates have important implications for the costs of alternative health insurance contracts, and hence for the optimal design of private insurance contracts or social insurance programs.

One aspect of this literature that we find remarkable is the near consensus on the nature of the endeavor: the attempt to quantify the response of medical spending with respect to its (out-of-pocket) price to the consumer. Yet, health insurance contracts in the United States are highly

[1] We are grateful to David Molitor and James Wang for outstanding research assistance, and to Amitabh Chandra, Kate Ho, Joe Newhouse and numerous seminar participants for helpful comments and suggestions. The Alcoa portion of the data were provided as part of an ongoing service and research agreement between Alcoa, Inc. and Stanford, under which Stanford faculty, in collaboration with faculty and staff at Yale University, perform jointly agreed-upon ongoing and ad hoc research projects on workers' health, injury, disability, and health care, and Mark Cullen serves as Senior Medical Advisor for Alcoa, Inc. We gratefully acknowledge support from the NIA (R01 AG032449), the National Science Foundation Graduate Research Fellowship program (Aron-Dine), the National Science Foundation Grant SES-0643037 (Einav), the John D. and Catherine T. MacArthur Foundation Network on Socioeconomic Status and Health and Alcoa, Inc. (Cullen), and the U.S. Social Security Administration through grant #5 RRC08098400-04-00 to the National Bureau of Economic Research as part of the SSA Retirement Research Consortium. The findings and conclusions expressed are solely those of the authors and do not represent the views of SSA, any agency of the Federal Government, or the NBER.

non-linear. so trying to estimate the behavioral response to a single out-of-pocket price is. in most cases. not a well-posed exercise. as it begs the question "which price?". A typical private health insurance plan has a deductible. a coinsurance rate. and an out-of-pocket maximum (or "stop loss"). The consumer faces a price of 100% of medical expenditures until he has spent the deductible. at which point the marginal price falls sharply to the coinsurance rate (typically around 10-20%). and then falls to zero once out-of-pocket expenditures have reached the stop-loss amount. Public health insurance programs. such as Medicare. also involve non-linear schedules. including occasionally schedules in which the marginal price rises over some expenditure range and then falls again (as in the famous "doughnut hole" in Medicare Part D prescription drug coverage).

In the context of such non-linear budget sets, trying to characterize an insurance policy by a single price could produce very misleading inferences. For example, one cannot extrapolate from estimates of the effect of coinsurance on health spending to the effects of introducing a high-deductible health insurance plan without knowing how forward looking individuals are in their response to health insurance coverage. A completely myopic individual would respond to the introduction of a deductible as if his "price" has sharply increased to 100%, whereas a fully forward looking individual with annual health expenditures that are likely to exceed the new deductible would experience little change in the effective marginal price of care and therefore might not change his behavior much.[2] Understanding how medical spending responds to the design of health insurance contracts therefore requires that we understand how consumers account for the non-linear budget schedule they face in making their medical consumption decisions. A fully rational, forward-looking individual who is not liquidity constrained should recognize that the "spot" price applied to a particular claim is not relevant; this nominal price should not affect his consumption decisions. Rather, the decision regarding whether to undertake some medical care should be a function only of the end-of-year price.

In this paper. we therefore investigate whether and to what extent individuals respond to the expected end-of-year price. or "future price," of medical care. We do so in the context of employer-provided health insurance in the United States, which is the source of over 85% of private health insurance coverage. Assessing whether individuals respond to the future price is empirically challenging. which may explain why there has been relatively little work on this topic. The key empirical difficulty arises because the spot price and the future price often vary jointly. A low spending individual faces both a high spot price (because all his spending falls below the deductible) and a high expected end-of-year price (because he does not expect to hit the deductible). while the

[2] Indeed, once one accounts for the non-linear contract design, even characterizing which insurance contract would provide greater incentives to economize on medical spending becomes a complicated matter. Consider, for example, two plans with a coinsurance arm that is followed by an out-of-pocket maximum of $5,000. Imagine that Plan A has a 10% coinsurance rate and plan B has a 50% coinsurance rate. Which plan would induce less spending? The naive answer would be that Plan B is less generous and would therefore lead to lower medical utilization. Yet, the answer depends on the distribution of medical spending without insurance, as well as on how forward looking individuals are. For example, an individual who suffers a compound fracture early in the coverage period and spends $10,000 on a surgery would effectively obtain full insurance coverage for the rest of the year under Plan B. but would face a 10% coinsurance rate (with a remaining $4,000 stop loss) under Plan A. We would therefore expect this individual to have greater medical utilization under Plan B.

opposite is true for a high spending individual. Similarly. the types of variation that have most often been used to estimate the impact of health insurance on medical spending – such as variation in deductibles or coinsurance rates – will change the spot price and the future price jointly. This makes it challenging to identify whether individuals respond to the future price without a tightly specified model of expectation formation. which in turn raises concerns about the extent to which any elasticity estimates are driven by these modeling assumptions.

The primary empirical exercise in this paper addresses this challenge by identifying situations in which individuals face the same spot price for their consumption decision, but have substantially different expected end-of-year prices. The key insight behind our empirical strategy is that, as a result of certain institutional features of employer-provided health insurance in the United States, individuals who join the same deductible plan in different months of the year initially face the same spot price, but different expected end-of-year prices. Employer-provided health insurance resets every year, typically on January 1. When new employees join a firm in the middle of the year, they obtain coverage for the remainder of the year. While their premiums are pro-rated, deductible amounts are fixed at their annual level. As a result, all else equal, the expected end-of-year price is increasing with the join month over the calendar year; individuals who join a plan later in the year have fewer months to spend past the deductible.

We use this feature in order to test for forward looking behavior in the response to health insurance contracts. In other words, we test the null of completely myopic behavior, which we define as consumption decisions that depend only on the spot price. We do so by comparing initial medical utilization across individuals who join the same plan in different months of the year. If individuals are forward looking in their healthcare consumption decisions, an individual who joins the plan earlier in the calendar year should (initially) spend more than an otherwise identical individual who joins the same plan later in the calendar year. By contrast, if individuals are myopic, the initial spending of an individual who joins the plan earlier should be the same as the initial spending of the individual who joins the same plan later. To account for potential confounders, such as seasonality in healthcare spending, we use patterns of initial utilization by join month for individuals who join no-deductible plans, in which the future price hardly varies over the course of the year. To operationalize this strategy empirically, we draw on data from several large employers with information on their plan details as well as their employees' plan choices, demographics, and medical claims.

We note that individuals may fail to exhibit forward-looking behavior not only because they are myopic but also if they are liquidity constrained or lack an understanding of their future budget constraint. If we had failed to reject the null of completely myopic behavior. we would have been unable to distinguish which of these factors was behind our result. In practice, however, we reject the null and estimate that conditional on the spot price of medical care, individuals who face a higher future price consume statistically significantly less (initial) medical care. It therefore appears that individuals understand something about the nature of their dynamic budget constraint and make their healthcare consumption decisions with at least some attention to forward-looking considerations.

In the last section of the paper we attempt to move beyond testing the null of complete myopia and toward quantifying the extent of forward looking behavior. We estimate that a ten cent increase in the future price (for a dollar of medical spending) is associated with a 6 to 8 percent decline in initial medical utilization. This implies an elasticity of initial medical utilization with respect to the future price of -0.4 to -0.6. To provide an economic interpretation of this estimate, we develop a stylized dynamic model in which utilization behavior in response to medical shocks depends on both the underlying willingness to substitute between health and residual income and the degree of forward looking behavior. We draw on additional data from the RAND Health Insurance Experiment to calibrate the model, and use the calibrated model to assess the extent of forward looking behavior implied by our estimates of the response of initial medical utilization to the future price. On the spectrum between full myopia (individuals respond only to the spot price) and textbook forward looking behavior (individuals respond only to the future price), our calibration results generally suggest that individuals' behavior is much closer to the former. Nonetheless, we illustrate that the degree of forward looking behavior we find still has a substantial effect on the response of annual medical spending to health insurance contracts relative to the spending response that would be predicted under either completely myopic or completely forward looking behavior. Thus, failing to account for dynamic considerations can greatly alter the predicted impact of non-linear health insurance contracts on annual medical expenditures.

Our paper links to the large empirical literature that tries to estimate moral hazard in health insurance, or the price sensitivity of demand for medical care. As already mentioned, much of this literature tries to estimate a demand elasticity with respect to a single price,[3] although different studies consider a different "relevant" price to which individuals are assumed to respond. For example, the famous RAND elasticity of -0.2 is calculated assuming individuals respond only to the spot price (Manning et al., 1987; Keeler and Rolph, 1988; Zweifel and Manning, 2000), while more recent estimates have assumed that individuals respond only to the expected end-of-year price (Eichner, 1997) or to the actual (realized) end-of-year price (Eichner, 1998; Kowalski, 2010). Our findings highlight the importance of thinking about the entire budget set rather than about a single price; this point was emphasized in some of the early theoretical work on the impact of health insurance on health spending (Keeler, Newhouse, and Phelps, 1977; Ellis, 1986) but until recently has rarely been incorporated into empirical work. Several papers on the impact of health insurance on medical spending – Ellis (1986), Cardon and Hendel (2001), and more recently Kowalski (2011), Marsh (2011), and our own work (Einav et al., 2011) – explicitly account for the non-linear budget set, but a (fully forward-looking) behavioral model is assumed, rather than tested.[4]

[3] See Chandra, Gruber, and McKnight (2007) for a recent review of this literature and its estimates.

[4] Non-linear pricing schedules are not unique to health insurance. Indeed, a large literature, going back at least to Hausman (1985), develops methods that address the difficulties that arise in modeling selection and utilization under non-linear budget sets, and applies these methods to other setting in which similar non-linearities are common, such as labor supply (Burtless and Hausman, 1978; Blundell and MaCurdy, 1999; Chetty et al., 2011), electricity utilization (Reiss and White, 2005), or cellular phones (Grubb and Osborne, 2009; Yao et al., 2011).

Outside of the context of health insurance, a handful of papers address the question of whether individuals respond at all to the non-linearities in their budget set, and which single price may best approximate the non-linear schedule to which individuals respond. This is the focus of Liebman and Zeckhauser (2004), Feldman and Katuscak (2006), and Saez (2010) in the context of the response of labor supply to the progressive income tax schedule, and of Borenstein (2009) and Ito (2010) in the context of residential electricity utilization. In most of these other contexts, as well as in our own previous work on moral hazard in health insurance (Einav et al., 2011), the analysis of demand in the presence of a non-linear pricing schedule is static. This is partly because in most non-health contexts information about intermediate utilization levels (within the billing or tax cycle) is not easy to obtain (for both consumers and researchers) and partly because dynamic modeling often introduces unnecessary complications in the analysis. In this sense, our current study – utilizing the precise timing of medical utilization within the contract year – is virtually unique within this literature in its explicit focus on the dynamic aspect of medical utilization, and its explicit account of expectation formation.[5]

Forward looking decision making plays a key role in many economic problems, and interest in the extent of forward looking behavior is therefore quite general. From this perspective, a closely related work to ours is Chevalier and Goolsbee's (2009) investigation of whether durable goods consumers are forward looking in their demand for college textbooks (they find that they are). Despite the obvious difference in context, their empirical strategy is similar to ours. They use the fact that static, spot incentives remain roughly constant (as pricing of textbook editions doesn't change much until the arrival of new editions), while dynamic incentives (the expected time until a new edition is released) change. A slightly cleaner aspect of our setting is that the constant spot prices and varying dynamic incentives are explicitly stipulated in the coverage contract rather than empirical facts that need to be estimated from data.

The rest of the paper proceeds as follows. Section 2 sketches a simple, stylized model of medical care utilization that is designed to provide intuition for the key concepts and our empirical strategy; the model serves as both a guide to some of our subsequent empirical choices, and as a framework that we use to benchmark the extent of forward looking behavior we estimate. In Section 3 we test for forward looking behavior. We start by describing the basic idea and the data we obtained to implement it, and then present the results. In Section 4 we calibrate the model from Section 2 to try to quantify the extent to which individuals are forward looking. Section 5 concludes.

2.2 A simple model

Consider a model of a risk-neutral forward-looking individual who faces uncertain medical expenditure, and is covered by a contract of (discrete) length T and deductible D.[6] That is, the individual

[5] An exception in this regard is Keeler and Rolph (1988), who, like us, test for forward looking behavior in health insurance contracts (but use a different empirical strategy and reach a different conclusion).

[6] Assuming risk neutrality in the context of an insurance market may appear an odd modeling choice. Yet, it makes the model simpler and more tractable and makes no difference for any of the qualitative insights we derive

pays all his expenditures out of pocket up to the deductible level D, but any additional expenditure is fully covered by the insurance provider.

The individual's utility is linear and additive in health and residual income, and we assume that medical events that are not treated are cumulative and additively separable in their effect on health. Medical events are given by a pair (θ, ω), where $\theta > 0$ denotes the total expenditure (paid by either the individual or his insurance provider) required to treat the event, and $\omega > 0$ denotes the (monetized) health consequences of the event if left untreated. We assume that individuals need to make a discrete choice whether to fully treat an event or not; events cannot be partially treated. We also assume that treated events are "fully" cured, and do not carry any other health consequences. Thus, conditional on an event (θ, ω), the individual's flow utility is given by

$$u(\theta, \omega; d) = \begin{cases} -min\{\theta, d\} & if\ treated \\ -\omega & if\ not\ treated \end{cases} \tag{2.1}$$

where $min\{\theta, d\}$ is the out-of-pocket cost associated with expenditure level θ, which is a function of d, the amount left to satisfy the deductible.

Medical shocks arrive with a per-period probability λ, and when they arrive they are drawn independently from a distribution $G(\theta, \omega)$. Given this setting, the only choice individuals make is whether to treat or not treat each realized medical event. Optimal behavior can be characterized by a simple finite horizon dynamic problem. The two state variables are the time left until the end of the coverage period which we denote by t, and the amount left until the entire deductible is spent which we denote by d. The value function $v(d, t)$ represents the present discounted value of expected utility along the optimal treatment path. Specifically, the value function is given by the solution to the following Bellman equation:

$$v(d,t) = (1-\lambda)\delta v(d,t-1) + \lambda \int max \begin{Bmatrix} -min\{\theta, d\} + \delta v(max\{d-\theta, 0\}, t-1), \\ -\omega + \delta v(d, t-1) \end{Bmatrix} dG(\theta, \omega), \tag{2.2}$$

with terminal conditions of $v(d, 0) = 0$ for all d. If a medical event arrives, the individual treats the event if the value from treating, $-min\{\theta, d\} + \delta v(max\{d - \theta, 0\}, t - 1)$, exceeds the value obtained from not treating, $-\omega + \delta v(d, t - 1)$.

The model implies simple and intuitive comparative statics: the treatment of a medical event is more likely when the time left on the contract, t, is higher and the amount left until the deductible is spent, d, is lower. This setting nests a range of possible behaviors. For example, "fully" myopic individuals ($\delta = 0$) would not treat any shock as long as the immediate negative health consequences of the untreated shock, ω, are less than the immediate out-of-pocket expenditure costs associated with treating that shock, $\min\{\theta, d\}$. Thus, if $\theta < d$, fully myopic individuals ($\delta = 0$) will not treat if $\omega < \theta$. By contrast, "fully" forward looking individuals ($\delta \approx 1$) will not treat shocks if the adverse health consequences, ω, are less than the expected end-of-year cost of treating this illness, which is

from the model.

given by $fp \cdot \theta$, where fp (for "future price") denotes the expected end-of-year price of medical care, which is the relevant price for a "fully" forward looking individual in deciding whether to consume care today. Thus, if $\theta < d$, fully forward looking individuals will not treat if $\omega < fp \cdot \theta$. That is, while fully myopic individuals consider the current, "spot", or nominal price of care (which in our example is equal to one), fully forward looking individuals only care about the future price.

To illustrate the implications of the model that will serve as the basis of our empirical strategy, we solve the model for a simple case, where we assume that $\lambda = 0.2$ and that medical events are drawn uniformly from a two-point support of $(\theta = 50, \omega = 50)$ and $(\theta = 50, \omega = 45)$. We use two different deductible levels (of 600 and 800) and up to 52 periods (weeks) of coverage. Figure 1 presents some of the model's implications for the case of $\delta = 1$. It uses metrics that are analogous to the empirical objects we later use in the empirical exercise. The top panel presents the expected end-of-year price of the individual as we change the deductible level and the coverage horizon. The expected end-of-year price in this example is $1 - \Pr(hit)$, where $\Pr(hit)$ is the fraction of individuals who hit the deductible by the end of the year. Individuals are, of course, more likely to hit the deductible as they have more time to do so or as the deductible level is lower. This ex-ante probability of hitting the deductible determines the individual's expectations about his end-of-year price. This future price in turn affects a forward looking individual's willingness to treat medical events. The bottom panel of Figure 1 presents the (cumulative) expected spending over the initial three months (12 weeks). Given the specific choice of parameter values, expected spending over the initial 12 periods is at least 60 (due to the per-period 0.1 probability of a medical event $(\theta = 50, \omega = 50)$ that would always be treated) and at most 120 (if all medical events are treated).

The key comparative static that is illustrated by Figure 1 – and that will form the basis of our empirical work – is how the expected end-of-year price (and hence initial spending by a forward looking individual) varies with the coverage horizon. For a given deductible, the expected end-of-year price is increasing as the coverage horizon declines (top panel of Figure 1) and therefore, for a forward looking individual, expected *initial* spending also declines as the coverage horizon declines (bottom panel of Figure 1). Specifically, when the coverage horizon is long enough and the deductible level low enough, forward looking individuals expect to eventually hit the deductible and therefore treat all events, so expected spending is 120. However, as the horizon gets shorter there is a greater possibility that the deductible would not get exhausted by the end of the year, so the end-of-year price could be 1 (rather than zero), thus making forward looking individuals not treat the less severe medical events of $(\theta = 50, \omega = 45)$.

The graphs also illustrate how the spot price of current medical care misses a great deal of the incentives faced by a forward looking individual. In the bottom panel of Figure 1 we see a fully forward looking individual's initial medical utilization (i.e., spending in the first 12 weeks) varying greatly with the coverage horizon despite a spot price that is always one. By contrast, for the cases we consider, a fully myopic individual $(\delta = 0)$ who only responds to the spot price has expected 12-week spending of 60, regardless of the coverage horizon t (see bottom panel).[7] Likewise, the expected three-month spending of individuals in a no-deductible plan does not vary with the coverage horizon, regardless of their δ, since the expected end-of-year price does not vary with the coverage horizon.

Finally, we note that while we have referred to δ as a measure of how "forward looking" the individual is, in practice a variety of different factors can push δ below 1 and induce a behavioral response to the current, "spot" price. These factors include not only myopia but also liquidity constraints (e.g., Adams, Einav, and Levin, 2009) and salience (e.g., Chetty and Saez, 2009; Liebman and Luttmer, 2011). Our research strategy does not distinguish between these, nor is it necessary to do so for predicting how spending will respond to changes in a non-linear budget set. However, these different sources that may affect behavior can be important for forecasting the effects of alternative public policy interventions or for extrapolating our results to alternative populations. We return to these issues briefly in the conclusion.

2.3 Testing for forward looking behavior

2.3.1 Basic idea

To test whether individuals exhibit forward looking behavior in their behavioral response to their health insurance contract, we design a test for whether individuals respond to the future price of medical consumption in a setting in which similar individuals face the same spot price (i.e., the nominal price at the time they make their medical consumption decision) but different future prices. In such a situation, we can test whether medical utilization changes with the future price, holding the spot price fixed, and interpret a non-zero coefficient as evidence of forward looking behavior and as a rejection of the null of complete myopia.

The central empirical challenge therefore is to identify individuals who face the same spot price but different future prices for medical consumption. Our novel observation is that the institutional features of employer-provided health insurance in the United States provide such variation. Specifically, we use the fact that unlike other lines of private insurance (e.g., auto insurance or home insurance), the coverage period of employer-provided health insurance is not customized to individual employees. This presumably reflects the need for synchronization within the company, such as benefits sessions, open enrollment periods, and tax treatment. Therefore, (annual) coverage begins (and ends, unless it is terminated due to job separation) at the same date – typically on January 1 – for almost all employees. Although all employees can choose to join a new plan for the subsequent year during the open enrollment period (typically in October or November), there are only two reasons employees can join a plan in the middle of the year: either they are new hires or they have

[7] A fully myopic individual ($\delta = 0$) would (like the fully forward looking individual) always treat ($\theta = 50, \omega = 50$) shocks but as long as he is still in the deductible range would never treat ($\theta = 50, \omega = 45$) shocks. Given this behavior, with a 600 or 800 deductible, there is a zero probability that the deductible would be reached within the first 12 weeks.

a qualifying event that allows them to change plans in the middle of the year.[8] In order to transition new employees (and occasionally existing employees who have a qualifying event) into the regular cycle, the common practice is to let employees choose from the regular menu of coverage options, to pro-rate linearly the annual premium associated with their choices, but to maintain constant (at its *annual* level) the deductible amount. As a result, individuals who are hired at different points in the year, but are covered by the same (deductible) plan, face the same spot price (of one) but different future prices. Thus, as long as employees join the company at different times for reasons that are exogenous to their medical utilization behavior, variation in hire date (or in the timing of qualifying events) generates quasi-experimental variation in the future price that allows us to test for forward looking behavior.

To illustrate, consider two identical employees who select a plan with an $800 (annual) deductible. The first individual is hired by the company in January and the second in July. The difference in their incentives is analogous to the simple model presented in Figure 1. Individuals who join in a later month during the year have a shorter coverage horizon t until coverage resets (on January 1). Individuals who join early in the year have a longer coverage horizon. The early joiners are therefore more likely to hit their deductible by the time their coverage resets. Therefore, as in the top panel of Figure 1, early joiners have a lower expected end-of-year price. As in the bottom panel of Figure 1, if individuals are forward looking, then early joiners have a greater incentive to utilize medical care upon joining the plan. Crucially, just after they get hired, both January and July joiners have yet to hit their deductible, so their spot price is (at least initially) the same. Thus, differences in (initial) spending cannot be attributed to differences in spot prices, and therefore must reflect dynamic considerations. By contrast, as Figure 1 also illustrates, if individuals are completely myopic (or join a plan with no deductible so that the expected end-of-year price does not vary with the month they join the plan), initial utilization will not vary for the early and later joiners.

2.3.2 Data

Data construction With this strategy in mind, we obtained claim-level data on employer-provided health insurance in the United States. We limited our sample to firms that offered at least one plan with a deductible (which would generate variation in expected end-of-year price based on the employee's join month, as in the top panel of Figure 1) and at least one plan with no deductible. The relationship between initial utilization and join month in the no-deductible plan is used to try to control for other potential confounding patterns in initial medical utilization by join month (such as seasonal flu); in a typical no-deductible plan, the expected end-of-year price is roughly constant by join month, so – absent confounding effects that vary by join month – initial medical utilization of employees covered by a no-deductible plan should not systematically vary with join month (bottom panel of Figure 1).

[8] Qualifying events include marriage, divorce, birth or adoption of a child, a spouse's loss of employment, or death of a dependent.

The data come from two sources. The first is Alcoa, Inc., a large multinational producer of aluminum and related products. We have four years of data (2004-2007) on the health insurance options, choices, and medical insurance claims of its employees (and any insured dependents) in the United States. We study the two most common health insurance plans at Alcoa, one with a deductible for in-network expenditure of $250 for single coverage ($500 for family coverage), and one with no deductible associated with in-network spending. While Alcoa employed (and the data cover) about 45,000 U.S.-based individuals every year, the key variation we use in this paper is driven by mid-year plan enrollment by individuals not previously insured by the firm, thus restricting our analysis to only about 7,000 unique employees (over the four years) that meet our sample criteria.[9] Of the employees at Alcoa who join a plan mid-year and did not previously have insurance at Alcoa that year, about 80% are new hires, while the other 20% are employees who were at Alcoa but uninsured at the firm, had a qualifying event that allowed them to change plans in the middle of the year, and chose to switch to Alcoa-provided insurance.

The Alcoa data are almost ideal for our purposes, with the important exception of sample size. Ex ante, sample size was a key concern given the large variation in medical spending across individuals. To increase statistical power we examined the set of firms (and plans) available through the National Bureau of Economic Research's (NBER) files of Medstat's MarketScan database. The data on plan choices and medical spending are virtually identical in nature and structure across the three firms (indeed, Alcoa administers its health insurance claims via Medstat); they include coverage and claim-level information from an employer-provided health insurance context, provided by a set of (anonymous) large employers.

We selected two firms that satisfied our basic criteria of being relatively large and offering both deductible and no-deductible options to their employees. Each firm has about 60,000 employees who join one of these plans in the middle of the year over the approximately six years of our data. This substantially larger sample size is a critical advantage over the Alcoa data. The disadvantages of these data are that we cannot tell apart new hires from existing employees who are new to the firm's health coverage (presumably due to qualifying events that allow them to join a health insurance plan in the middle of the year), we cannot distinguish between in-network and out-of-network spending, there is less demographic information on the employees, and the coinsurance rate for one of the plans in one of the firms is not known.

Because employers in MarketScan are anonymous (and we essentially know nothing about them), we will refer to these two additional employers as firm B and firm C. We focus on two plans offered by firm B. We have five years of data (2001-2005) for these plans, during which firm B offered one plan with no in-network deductible and one plan that had a $150 ($300) in-network single (family) deductible. The data for firm C are similar, except that the features of the deductible plan have changed slightly over time. We have seven years of data for firm C (1999-2005), during which the firm continuously offered a no-deductible plan (in-network) alongside a plan with a deductible. The

[9] We restrict our analysis to employees who are not insured at the firm prior to joining a plan in the middle of the year because if individuals change plans within the firm (due to a qualifying event), the deductible would not reset.

deductible amount increased over time. with a single (family) in-network deductible of \$200 (\$500) during 1999 and 2000. of \$250 (\$625) during 2001 and 2002. and \$300 (\$750) during 2004 and 2005.

Table 1 summarizes the key features of the plans (and their enrollment) that are covered by our final data set. In all three firms. we limit our analysis to employees who join a plan between February and October. and who did not have insurance at the firm immediately prior to this join date. We omit employees who join in January for reasons related to the way the data are organized that make it difficult to tell apart new hires who join the firm in January from existing employees. We omit employees who join in November or December because, as we discuss in more detail below. we use data from the first three months after enrollment to construct our measures of "initial" medical utilization. Table 1 also summarizes, by plan, the limited demographic information we observe on each covered employee, namely the type of coverage they chose (family or single), and the employee's gender, age, and enrollment month.[10]

Measuring the expected end-of-year price Table 2 describes the key variation we use in our empirical analysis. For each plan, we report the expected end-of-year price as a function of the time within the year an employee joined the plan.[11] Specifically, we define the expected end-of-year price, or future price, fp, as

$$fp_{jm} = 1 - \Pr(hit_{jm}), \tag{2.3}$$

where $\Pr(hit_{jm})$ is the probability an employee who joins plan j in month m will hit (i.e., spend more than) the in-network deductible by the end of the year; we calculate $\Pr(hit)$ as the fraction of employees in a given plan and join month who have spent more than the in-network deductible by the end of the year.[12] For example, consider a plan with a \$500 deductible and full coverage for any medical expenditures beyond the deductible. If 80% of the employees who joined the plan in February have hit the deductible by the end of the year, the expected end-of-year price would be $0.8 \cdot 0 + 0.2 \cdot 1 = 0.2$. If only 40% of the employees who joined the plan in August have hit the deductible by the end of the year, their expected end-of-year price would be $0.4 \cdot 0 + 0.6 \cdot 1 = 0.6$. Thus, the future price is the average (out-of-pocket) end-of-year price of an extra dollar of in-network spending. It is a function of one's plan j, join month m, and the annual spending of all the employees in one's plan and join month.

Table 2 summarizes the average future price for each plan based on the quarter of the year in which one joins the plan. For plans with no deductible ($A0$, $B0$, and $C0$), the future price is mechanically zero (since everyone "hits" the zero deductible), regardless of the join month. For

[10] In each firm we lose roughly 15 to 30 percent of new plan joiners because of some combination of missing information about the employee's plan, missing plan details, or missing claims data (because the plan is an HMO or a partially or fully capitated POS plan).

[11] In this and all subsequent analyses we pool the three different deductible plans in firm C which were offered at different times over our sample period.

[12] We calculate $\Pr(hit)$ separately for employees with individual and family coverage (since both the deductible amount and spending patterns vary with the coverage tier), and therefore in all of our analyses fp varies with coverage tier. However, for conciseness, in the tables we pool coverage tiers and report the (weighted) average across coverage tiers within each plan.

deductible plans, however, the future price varies with the join month. Only a small fraction of the individuals who join plans late in the year (August through October) hit their deductible, so their future price is greater than 0.8 on average. In contrast, many more employees who join a deductible plan early in the year (February to April) hit their deductible, so for such employees the future price is just over 0.5. Thus, early joiners who select plans with a deductible face an average end-of-year price that is about 30 percentage points lower than the end-of-year price faced by late joiners. Yet, initially (just after they join) both types of employees have yet to hit their deductible, so they all face a spot price of one. Differences in initial spending between the groups therefore plausibly reflects their dynamic response to the future price. This baseline definition of the future price – the fraction of employees who join a given plan in a given month whose spending does not exceed the in-network deductible by the end of the calendar year – will be used as the key right hand variable in much of our subsequent empirical work.

Our baseline measure of the future price abstracts from several additional characteristics of the plans, which are summarized in Appendix Table A1. First, it ignores any coinsurance features of the plans. Plans A0, A1, and C1-C3 all have a 10% coinsurance rate, while plans B0 and C0 have a zero coinsurance rate. The coinsurance rate for plan B1 is unknown (to us). Second, we use only the in-network plan features and assume that all spending occurs in network. In practice, each plan (including the no-deductible plan) has deductibles and higher consumer coinsurance rates for medical spending that occurs out of network.

There are two consequences of these abstractions, both of which bias any estimated impact of the future price on behavior toward zero. First, abstracting from these features introduces measurement error into the future price. Second, our analysis assumes that for the no-deductible plans there is no variation in the future price for employees who join in different months (i.e., the spot price and the future price are always the same). In practice, both a positive in-network coinsurance rate (prior to the stop-loss) and the existence of out-of-network deductibles in all of the no-deductible (in-network) plans mean that the future price also increases with the join month for employees in the no-deductible plans. In the robustness section below we show that accounting for these additional features – to the extent we are able to – makes little quantitative difference to either our measurement of the future price or its estimated effect.

A final point worth noting about our definition of the future price is that it is constructed based on the observed spending patterns of people who join a specific plan (and coverage tier) in a specific month. For forward looking individuals, this spending may of course be influenced by the future price. As we discuss in more detail below, this is not a problem for testing the null of complete myopia (because under this null spending is not affected by the future price). Yet, for quantifying the extent of forward looking behavior in Section 4 we will implement an instrumental variable strategy designed to purge the calculated future price of any endogenous spending response.

2.3.3 Estimating equations and results

Patterns of initial utilization by plan and join month We proxy for "initial" utilization with two alternative measures. The first is a measure of the time (in days) to the first claim, while the second is a measure of total spending (in dollars) over some initial duration (we will use three months). In both cases, the measures of utilization encompass the utilization of the employee and any covered dependents.

Average three month spending in our sample is about \$600. It is zero for about 42% of the sample. Since time to first claim is censored at as low a value as 92 days (for individuals who join in October), we censor time to first claim at 92 for all the individuals (regardless of join month) who have their first claim more than 92 days after joining the firm's coverage. The average time to first claim for the remaining 58% of the individuals is 35 days, so with 42% of the sample censored at 92 days, the sample average for the censored variable is 58 days.

Table 3 reports summary statistics for these measures of initial medical utilization by join month for each plan. These statistics already indicate what appears to be a response to dynamic incentives. For the no-deductible plans the average initial spending (left panel) and time to first claim (right panel) are somewhat noisy, but do not reveal any systematic relationship with join month. By contrast, employees who are in deductible plans appear to spend substantially less within the first three months after joining the plan, or have a substantially longer time to first claim, if they join the plan later in the year, presumably due to dynamic considerations. As illustrated in the bottom panel of Figure 1, this is exactly the qualitative pattern one would expect from forward looking individuals.

We operationalize this analysis a little more formally by regressing the measures of initial utilization on join month. A unit of observation is an employee e who joins health insurance plan j during calendar month m. As mentioned, we limit attention to employees who join new plans between February and October, so $m \in \{2, ...10\}$. The simplest way by which we can implement our strategy is to look within a given health plan that has a positive deductible associated with it and regress a measure of initial medical utilization y_e on the join month m_e and possibly a set of controls x_e, so that:

$$y_e = \beta_j m_e + x_e' \gamma + u_e. \tag{2.4}$$

Absent any confounding influences of join month on y_e, we would expect an estimate of $\beta_j = 0$ for deductible plans if individuals are fully myopic ($\delta = 0$) and $\beta_j < 0$ for spending ($\beta_j > 0$ for time to first claim) if individuals are not ($\delta > 0$). We include an additional covariate for whether the employee has family (as opposed to single) coverage to account for the fact that the deductible varies within a plan by coverage tier (see Table 1) and that there naturally exist large differences in average medical utilization in family vs. single coverage plans.

For our analysis of initial spending, our baseline dependent variable is $\log(s + 1)$, where s is total medical spending (in dollars) by the employee and any covered dependents during their first three months in the plan. Given that medical utilization is highly skewed, the log transformation helps in improving precision and reducing the effect of outliers.[13] An added attraction of the log specification is that it facilitates comparison of the results to those from our analysis of time to first claim. For the latter analysis, we use a Tobit specification on log($time$), where $time$ measures the time to first claim (in days) by the employee and any covered dependents; the Tobit is used to account for the censoring at 92 days described above. We explore alternative functional forms for both dependent variables below.

Columns (1) and (3) of Table 4 report results from estimating equation 4 on these two dependent variables, separately for each plan. The key right-hand-side variable is the join month, enumerated from 2 (February) to 10 (October). In plans that have a deductible ($A1$, $B1$, and $C1$-$C3$), dynamic considerations would imply a negative relationship between join month and initial spending and a positive relationship between join month and time to first claim. The results show exactly this qualitative pattern.

Patterns of initial utilization by join month for deductible vs. no-deductible plan If seasonality in medical utilization is an important factor, it could confound the interpretation of the estimated relationship that we have just discussed as a test for the null of full myopia. For example, if spending in the spring is greater than spending in the summer due to, say, seasonal flu, then we may incorrectly attribute the decline in "spot" utilization for late joiners as a response to dynamic incentives. To address this concern (and other possible confounding differences across employees who join plans at different months of the year), we use as a control group employees within the same firm who join the health insurance plan with no deductible in different months. As discussed earlier, such employees are in a plan in which the spot price and future price are (roughly) the same so that changes in their initial utilization over the year (or lack thereof) provides a way to measure and control for factors that influence initial utilization by join month that are unrelated to dynamic incentives.

Columns (1) and (3) of Table 4, discussed earlier, also show the plan-level analysis of the relationship between initial medical utilization and join month for the no-deductible plans ($A0$, $B0$, and $C0$). The coefficient on join month for the no-deductible plans tends to be much smaller than the coefficient for the deductible plan in the same firm (and is often statistically indistinguishable from zero). This suggests that the difference-in-difference estimates of the pattern of spending by join month in deductible plans relative to the analogous pattern in no-deductible plans will look very similar to the patterns in the deductible plans. Indeed, this is what we find, as reported in columns (2) and (4) of Table 4, which report this difference-in-difference analysis in which the no-deductible plan (within the same firm) is used to control for the seasonal pattern of initial utilization by join month in the "absence" of dynamic incentives. Specifically, the difference-in-differences specification

[13] While conceptually a concave transformation is therefore useful, we have no theoretical guidance as to the "right" functional form; any transformation therefore (including the one we choose) is ad hoc, and we simply choose one that is convenient and easy to implement. We note however that Box-Cox analysis of the $s + 1$ variable suggests that a log transformation is appropriate.

is

$$y_e = \beta' m_e D_j + \mu_j + \tau_m + x_e' \gamma' + \upsilon_e, \tag{2.5}$$

where μ_j are plan fixed effects, τ_m are join-month fixed effects, and D_j is a dummy variable that is equal to one when j is a deductible plan. The "plan fixed effects" (the μ_j's) include separate fixed effects for each plan by coverage tier (family or single) since the coverage tier affects the deductible amount (see Table 1). Again, our coefficient of interest is β', where $\beta' = 0$ would be consistent with the lack of response to dynamic incentives (i.e., full myopia) while $\beta' < 0$ (for spending; or $\beta' > 0$ for time to first claim) implies that the evidence is consistent with forward looking behavior. Since we are now pooling results across plans (deductible and no-deductible plans), the parameter of interest β' no longer has a j subscript.

The results in Table 4 indicate that, except at Alcoa where we have much smaller sample sizes, the difference-in-difference estimates for each firm are all statistically significant and with the sign that is consistent with dynamic considerations. For example, in Firm B we find that enrollment a month later in a plan with a (\$150 or \$300) deductible relative to enrollment a month later in a plan with no deductible is associated with an 8% decline in medical expenditure during the first three months, and a 3% increase in the time to first claim. In Firm C these numbers are a 2% decline and a 2% increase, respectively.

Of course, employees who self select into a no-deductible plan are likely to be sicker and to utilize medical care more frequently than those employees who select plans with a deductible (due to both selection and moral hazard effects). Indeed, Table 1 shows that there are, not surprisingly, some observable differences between employees within a firm who choose the no-deductible option instead of the deductible option. Our key identifying assumption is that while initial medical utilization may differ on average between employees who join deductible plans and those who join no-deductible plans, the within-year pattern of initial utilization by join month does not vary based on whether the employee joined the deductible or no-deductible plan except for dynamic incentives. In other words, we assume that any differences in initial utilization between those who join the no-deductible plan and the deductible plan within a firm can be controlled for by a single (join month invariant) dummy variable. We return to this below, when we discuss possible threats to this identifying assumption and attempt to examine its validity.

Testing the relationship between expected end-of-year price and initial utilization In order to provide an economic interpretation to the parameter of interest, it is useful to convert the key right-hand-side variable, join month (m_e), into a variable that is closer to the underlying object of interest: the expected end-of-year price. We therefore start by analyzing variants of the single-plan analysis (equation 4) and the difference-in difference analysis (equation 5) in which we replace the join month variable (m_e) with the future price variable fp defined earlier (recall equation 3 for a definition, and Table 2 for summary statistics). The estimating equations are thus modified to

$$y_e = \widetilde{\beta}_j fp_m + x_e' \widetilde{\gamma} + \widetilde{u}_e, \tag{2.6}$$

and

$$y_e = \widetilde{\beta}' fp_{jm} + \widetilde{\mu}_j + \widetilde{\tau}_m + x_e'\widetilde{\gamma}' + \widetilde{v}_e, \tag{2.7}$$

where (as before) $\widetilde{\mu}_j$are plan (by coverage tier) fixed effects, and $\widetilde{\tau}_m$ are join-month fixed effects. This transformation also aids in addressing the likely non-linear effect of join month on both expected end-of-year price and on expected spending. Figure 1 illustrates how this relationship may be non-linear, and Table 2 indicates that, indeed, our measure of the end-of-year price varies non-linearly over time.

Table 5 reports the results. The first three rows report the results for each firm. We report the results for the deductible plan in each firm in columns (1) and (3) and the difference-in-difference results that use the deductible and no-deductible plan within each firm in columns (2) and (4).[14] The difference-in-difference results in Firm B and Firm C (where the sample sizes are much bigger) suggest that a 10 cent increase in the expected end-of-year price is associated with an 8 to 17 percent reduction in initial medical spending and with a 2.5 to 7 percent increase in the time to first claim. These results are almost always statistically significant.

Thus far, all of the analysis has been of single plans or pairs of plans within a firm. The use of future price (rather than join month) also allows us to more sensibly pool results across firms and summarize them with a single number, since the relationship between join month and future price will vary both with the level of the deductible (see Figure 1) and with the employee population. In pooling the data, however, we continue to rely on only within firm variation, since we know little about the different firms or about how comparable (or not) their employee populations are (although we show in the appendix that in practice this does not make a substantive difference to the results). Thus, our final specification allows the join month dummy variables $\widetilde{\tau}_m$'s to vary by firm, so that all of the identification is coming from the differential effect of the join month on employees in deductible plans relative to no-deductible plans within the same firm. That is, we estimate

$$y_e = \widetilde{\widetilde{\beta}}' fp_{jm} + \widetilde{\widetilde{\mu}}_j + \widetilde{\widetilde{\tau}}_{mf} + x_e'\widetilde{\widetilde{\gamma}}' + \widetilde{\widetilde{v}}_e, \tag{2.8}$$

where $\widetilde{\widetilde{\tau}}_{mf}$ denotes a full set of join month by firm fixed effects. The bottom rows of Table 5 report the results from this regression. The OLS results (penultimate row of Table 5) will represent our baseline specification in the rest of this section. We defer discussion of the IV results (last row of Table 5) to the next section.

The effect of future price is statistically significant for both dependent variables. The OLS results in the penultimate row indicate that an increase of 10 cents in the future price is associated with an 11% decline in initial medical spending and a 3.6% increase in time to first claim. Overall, the results suggest that we can reject the null of complete myopia ($\delta = 0$). Individuals appear to respond to the future price of medical care in making current medical care utilization decisions. In other words, among individuals who face the same spot price of medical care, individuals who

[14] Note that, by design, fp is constant for no-deductible plans, so that we cannot estimate the single-plan analysis of the relationship between initial medical utilization and future price for the no-deductible plans.

face a higher expected end-of-year price – because they join the plan later in the year – initially consume less medical care.

We also investigated the margin on which the individual's response to the future price occurs. About three quarters of medical expenditures in our data represent outpatient spending; per episode. inpatient care is more expensive and perhaps less discretionary than outpatient care. Perhaps not surprisingly therefore, we find clear evidence of a response of outpatient spending to the future price, but we are unable to reject the null of no response of inpatient spending to the future price (although the point estimates are of the expected sign); Appendix Table A2 contains the results.

Robustness We explored the robustness of our results to a number of our modeling choices. The first six rows of Table 6 shows that our finding is quite robust across alternative functional forms for the dependent variable. The first row shows the baseline results, where for initial spending the dependent variable is $\log(s+1)$, where s is total medical spending in the three months after joining the plan, and for time-to-first-claim we estimate a Tobit model for $\log(time)$, where $time$ is the number of days until the first claim, censored at 92.

Row (2) of Table 6 uses levels (rather than logs) of s and $time$ (maintaining the Tobit specification for the $time$ analysis). The statistically significant estimates are comparable in magnitude to those in the baseline specification. Relative to the mean of the dependent variable, the results in row (2) suggest that a 10 cent increase in the future price is associated with a 7% decline in initial spending (compared to an 11% decline estimated in the baseline specification), and a 2.5% increase in the time to first claim (compared to a 3.6% increase in the baseline). In row (3) we report results from quasi-maximum likelihood Poisson estimation and calculate the fully-robust variance covariance matrix (Wooldridge, 2002, pp. 674-676); this is an alternative proportional model to the log specification, and one that allows us to incorporate the frequent occurrence of zero initial spending without adding 1 when the dependent variable is based on three-month spending. The estimate is still statistically significant, although somewhat smaller than our baseline estimate for initial spending (suggesting that a 10 cent increase in the future price is associated with a 7% rather than 11% decline in initial spending).

The next three rows investigate alternative ways of handling the time to first claim analysis. Row (4) shows that estimating the baseline specification by OLS instead of Tobit produces estimates that are still statistically significant but are somewhat smaller than the baseline (a 1.1% rather than 3.6% increase in time to first claim). Row (5) reports result from estimating a censored-normal regression on our baseline dependent variable $\log(time)$, which allows for the censoring value to vary across observations. This allows us to make use of the fact that while we only observe 92 days of medical claims for individuals who join in October, we can expand the observation period for individuals who join in earlier months. The advantage of such a specification is that it makes use of more information; the disadvantage is that it may not be as comparable to the spending estimates since it implicitly gives more weight to individuals who join earlier in the year. The results are virtually identical to the baseline specification. In row (6) we estimate a Cox semi-parametric proportional hazard model of the time to first claim (censored at 92 days for all observations).[15] Consistent with the previous specifications, the results indicate that an increase in the future price is associated with a statistically significant decline in the probability of a claim arrival (i.e., a longer time to first claim).

In Appendix Table A3 (Panels A and B) we further show the robustness of our results to alternative choices of covariates regarding the firm and coverage tier fixed effects. We also explore an alternative measure of the future price which, unlike our baseline measure, accounts for the in-network coinsurance rates in both the deductible and no-deductible plans for the two firms in which this information is available (Alcoa and Firm C; see Appendix Table A1). Accounting for the in-network coinsurance rates for Alcoa and Firm C makes little difference to either our measurement of the future price (Appendix Table A4) or its estimated effect (Appendix Table A3, Panel C), although the results in Appendix Table A3 suggest that, as expected (see discussion in Section 3.2), not accounting for the coinsurance rate slightly biases downward the estimated impact of the future price in our baseline specification.[16]

2.3.4 Assessing the identifying assumption

The results suggest that we can reject the null of complete myopia in favor of some form of forward looking behavior. The key identifying assumption behind this interpretation of the results is that there are no confounding differences in initial medical utilization among employees by their plan and join month. In other words, any differential patterns of initial medical utilization that we observe across plans by join month is caused by differences in expected end-of-year price. This identifying assumption might not be correct if for some reason individuals who join a particular plan in different months vary in their underlying propensity to use medical care. In particular, one might be concerned that the same forward looking behavior that may lead to differential medical care utilization might also lead to differential selection into a deductible compared to a no-deductible plan on the basis of join month.

A priori, it is not clear if forward looking individuals would engage in differential selection into a deductible vs. no-deductible plan based on the month they are joining the plan. A selection story that would be most detrimental to the interpretation of our results is that individuals who have high expected initial medical expenditure would be more likely to select the no-deductible plan later in the year. For example, if an individual knows that all he needs is a single (urgent)

[15] Van der Berg (2001) discusses the trade-offs involved in analyzing a duration model using a linear model with the logarithm of the duration as the dependent variable, relative to a proportional hazard model. As he explains, neither model strictly dominates the other.

[16] We do not observe the breakdown of spending by in-network vs. out-of-network in Firm B or Firm C, so we cannot account for out-of-network spending in our construction of the future price at either of these firms. We do know that in Alcoa, where the data allow us to tell apart in-network spending from out-of-network spending, about 95% of the spending is done in network. We therefore suspect that the accounting for out-of-network spending and out-of-network features of the plan would have little quantitative impact on our estimates of either the future price or the response to it.

doctor's appointment of \$100 (which is below the deductible amount). it may be worth joining the no-deductible plan (and paying the higher monthly premium) if he joins the plan later in the year but not earlier in the year. as late in the year the incremental premium of a no-deductible plan is lower and would be less than the \$100 benefits it would provide. This would introduce a positive relationship between individuals who join the no-deductible plan in later months and initial medical utilization and could cause us to erroneously interpret the lack of such a pattern in the deductible plans as evidence that individuals respond to the future price.

On the other hand. there are many reasons to expect no selection. even in the context of forward looking individuals. if there are additional choice or preference parameters governing insurance plan selection that do not directly determine medical utilization. For example. if individuals anticipate the apparently large switching costs associated with subsequent re-optimization of plan choice (as in Handel. 2011) they might make their initial. mid-year plan choice based on their subsequent optimal open enrollment selection for a complete year. In such a case. we would not expect differential selection into plans by join month. Ultimately. whether there is quantitatively important differential selection and its nature is an empirical question.

The summary statistics in Table 2 present some suggestive evidence that individuals may be (slightly) more likely to select the deductible plan relative to the no-deductible plan later in the year.[17] Quantitatively. however. the probability of selecting the deductible vs. no-deductible plan is very similar over the course of the year. When we regress an indicator variable for whether the employee chose a deductible plan on the employee's join month (enumerated, as before, from 2 to 10), together with a dummy variable for coverage tier and firm fixed effects to parallel our main specification. the coefficient on join month is 0.0034 (standard error 0.0018). Qualitatively. the pattern of greater probability of choosing a deductible later in the year is the opposite of what could produce a confounding explanation for our main empirical results. More importantly. quantitatively the results suggest trivial differential plan selection by join month; joining a month later is associated with only a 0.3 percentage point increase in the probability of choosing the deductible plan. or 0.9% relative to the 32% probability of choosing the deductible plan in the sample. This very similar share of choices of deductible vs. no-deductible plans over the course of the year implies that differential plan selection is unlikely to drive our findings.

We also examined whether the observable characteristics – i.e. age and gender – of individuals joining a deductible vs. no-deductible plan within each of the three firms varied by join month. In general. the results (shown in Appendix Table A5) show little evidence of systematic differences by join month.[18] To examine whether our findings are sensitive to these observable differences. in Row 7 of Table 6 we re-estimate our baseline specification (equation 8) adding controls for the

[17] Over the three join quarters shown in Table 2. the share joining the deductible plan varies in Alcoa from 0.49 to 0.53 to 0.53, in firm B from 0.20 to 0.22 to 0.19. and in firm C from 0.38 to 0.40 to 0.44.

[18] While there are two exceptions that show statistically significant differential selection by join month. they are both quantitatively trivial. Employees at Alcoa who join a deductible vs. no-deductible plan one month later in the year are 0.9 percentage points (about 2%) more likely to be female. Employees at Firm B who join a deductible vs. no-deductible plan one month later in the year are 0.6 percentage points (about 2%) less likely to be over 45 (or 0.2 months younger (not shown in the table)).

observable demographic characteristics of the employees: employee age, gender, and join year (see Table 1). In keeping with the "within-firm" spirit of the entire analysis, we interact each of these variables with the firm fixed effects. This specification controls for potential observable differences across employees within a firm by plan type and join month. The results indicate that the impact of these controls is neither substantial nor systematic. The effect of a 10 cent increase in the expected end-of-year price on initial spending declines from 11% in the baseline specification to 10% with the demographic controls, while the effect on time to first claim increases from 3.6% in the baseline specification to 5.2% with the demographic controls. All the results remain statistically significant.

As another potential way to investigate the validity of the identifying assumption, we implement an imperfect "placebo test" by re-estimating our baseline specification (equation 8 with the dependent variable as the "initial" medical utilization in the second year the employee is in the plan. In other words, we continue to define "initial medical utilization" relative to the join month (so that the calendar month in which we measure initial medical utilization varies in the same way as in our baseline specification across employees by join month) but we now measure it in the second year the employee is in the plan. For example, for employees who joined the plan in July 2004, we look at their medical spending during July through September 2005. In principle, when employees are in the plan for a full year there should be no effect of join month (of the previous year) on their expected end-of-year price, and therefore no difference in "initial" utilization by join month across the plans. In practice, the test suffers from the problem that the amount of medical care consumed in the join year could influence (either positively or negatively) the amount consumed in the second year, either because of inter-temporal substitution (which could generate negative serial correlation) or because medical care creates demand for further care (e.g., follow up visits or further tests), which could generate positive serial correlation.

Row (8) of Table 6 shows the baseline results limited to the approximately 60% of the employees who remain at the firm for the entire second year. We continue to find the same basic pattern in this smaller sample although the point estimate declines (in absolute value) and the time to first claim results are no longer statistically significantly different from zero. For this subsample of employees, row (9) shows the results when we now measure "initial medical spending" in the same three months but in the second year.[19] Here we find that an increase in the future price is associated with a much smaller and statistically insignificant decline in medical spending measured over the same three month period but in the second year. We interpret this as generally supportive of the identifying assumption, and suggestive of positive serial correlation in medical spending.

Finally, in row (10) we investigate the extent to which the decrease in utilization in response to a higher future price represents inter-temporal substitution of medical spending to the next year. Such inter-temporal substitution would not be a threat to our empirical design – indeed, it might be viewed as evidence of another form of forward-looking behavior in medical spending – but it would affect the interpretation of our estimates and is of interest in its own right. We therefore

[19] We perform this "second year" analysis only for the dependent variable "initial medical spending" as it seemed awkward to us to examine "time to first claim" from an arbitrary starting point in the second year (when in fact the individual has had prior months to make his first claim).

re-run our baseline specification but now with the dependent variables measured in January to March of the second year. The results indicate that individuals who face a higher future price (and therefore consume less medical care) also consume less medical care in the beginning of the subsequent year. This suggests that inter-temporal substitution, in the form of postponement of care to the subsequent calendar year, is unlikely to drive the decrease in care associated with a higher future price.

2.4 Quantifying forward looking behavior

Our results thus far have rejected the null of no response to the future price and presented evidence consistent with some form of forward looking behavior. A natural subsequent question is to ask how forward looking the behavior is. In other words, having rejected one extreme of complete myopia ($\delta = 0$), we would like to get a sense of where individuals lie on the spectrum between complete myopia ($\delta = 0$) and "full" forward looking behavior ($\delta \approx 1$). Relatedly, we are also interested in the implications (relative to either of these extremes) of the amount of forward looking behavior we detect for the the impact of alternative health insurance contracts on annual medical spending.

2.4.1 Quantifying the effect of the future price on initial medical utilization

We start by quantifying the elasticity of initial medical utilization with respect to the future price. The results reported in the previous section tested whether there was a relationship between the future price and initial medical utilization. However, a concern with interpreting this relationship as the causal effect of the future price on initial medical utilization is that there is a mechanical relationship between initial medical utilization (the dependent variable) and our measure of the future price (the right-hand-side variable). The future price is a function of the plan (by coverage tier) chosen, the month joined, and the monthly medical spending of people who join that plan (by coverage tier) in that month; thus, the future price is a function of medical spending which is also used in constructing the dependent variable. This is not a concern for testing the null of complete myopia (i.e., testing whether the coefficient on the future price is zero) which was the focus of the last section, because under the null of complete myopia medical spending is not a function of the future price. However, under the alternative hypothesis that individuals are forward looking, this can bias away from zero the estimated response to the future price.

To address this concern, we present results from estimating an instrumental variable version of equation 8 in which we instrument for the future price with a simulated future price. Like the future price, the simulated future price is computed based on the characteristics of the plan (by coverage tier) chosen and the month joined. However, unlike the future price which is calculated based on the spending of people who joined that plan (by coverage tier) that month, the simulated future price is calculated based on the spending of all employees in that firm and coverage tier in our sample who joined either the deductible or no-deductible plan, regardless of join month.[20] By

[20] Specifically, for every employee in our sample in a given firm and coverage tier (regardless of plan and join month)

using a *common sample* of employee spending that does not vary with plan or join month, the instrument is "purged" of any potential variation in initial medical utilization that is correlated with plan and join month, in very much the same spirit as Currie and Gruber's (1996) simulated Medicaid eligibility instrument. An additional attraction of this IV strategy is that it helps correct for any measurement error in our calculated future price (which would bias the coefficient toward zero). On net, therefore, the OLS may be biased upward or downward relative to the IV.

The bottom row of Table 5 shows the results from this IV strategy. As would be expected, the first stage is very strong and the IV estimates are statistically significant.[21] For the dependent variable log initial spending, the point estimate from the IV results suggests that a 10 cent increase in the expected end-of-year price is associated with a 7.8% decline in initial medical spending. Given an average expected end-of-year price for people in our sample who choose the deductible plan of about 70 cents, this suggests an elasticity of initial medical utilization with respect to the future price of -0.56. For the dependent variable log time to first claim, the IV results suggest that a 10 cent increase in the expected end-of-year price is associated with a 5.6% increase in the time to first claim, or an elasticity of initial medical utilization with respect to the future price of about -0.39.[22]

2.4.2 Mapping the estimated elasticity to economic primitives of interest

There are (at least) two related reasons why this estimate of the elasticity of initial medical utilization with respect to the future price is an unsatisfactory answer to the question: how important is forward looking behavior? The first reason is that this elasticity measures the effect of future price on *initial* spending, while we suspect that *total* (annual) spending is the outcome variable

we compute their monthly spending for all months that we observe them during the year that they join the plan, creating a common monthly spending pool. We then simulate the future price faced by an employee in a particular plan and join month by drawing (with replacement) 110,000 draws of monthly spending from this common pool, for every month we need a monthly spending measure. For the first month we draw from the pool of first month spending (since people may join the plan in the middle of the month, the first month's spending has a different distribution from other months) whereas for all other months in the plan that year we draw from the pool (across families and months) of non first month spending. For each simulation we then compute the expected end-of-year price based on the draws.

[21] The first stage coefficient from the regression of the future price on the instrument (as well as plan-by-coverage tier fixed effects and firm-by-start month fixed effects) yields a coefficient (on the instrument) of 0.56 (standard error 0.024).

[22] In principle, the IV estimate of the impact of the future price on the first three months' spending could be biased upward since, over the first three months, 17% of the individuals in deductible plans spend past the deductible. If individuals are at least partially forward looking, the probability of hitting the deductible in the first three months could be correlated with join month, which would introduce variation during the first three month in the spot price among individuals who join the same plan in different months. Once again, this is not a problem for testing the null of complete myopia; nor is it a problem when the dependent variable is the time to first claim (since the spot price is the same for all individuals within a plan at the time of first claim). In practice, moreover, any upward bias is likely unimportant quantitatively. We estimate a virtually identical response to the future price when the dependent variable is based on two-month (instead of three-month) spending, even though the fraction hitting the deductible within the initial utilization period (and therefore the likely magnitude of the bias) drops by almost a half. Moreover, there is no noticeable trend in the likelihood of hitting the deductible within the first three months by the join month; hitting the deductible within a short time after enrollment therefore appears to be primarily determined by large and possibly non-discretionary health shocks, rather than an endogenous spending response to the future price.

of interest for most research or policy questions associated with health insurance utilization. The importance of dynamic incentives for annual spending may well be much lower than for initial spending since the wedge between the spot price and the future price becomes smaller as health shocks accumulate within the year and/or the end of the coverage period nears.

The second reason is that it is difficult to assess whether the elasticity is large or small without appropriate benchmarks. We would like to compare our estimated elasticity with respect to the future price to the "primitive" price elasticity, i.e. the underlying elasticity that is driven by substitution between health and income and is purged of dynamic incentives. However, the same motivation that prompted us to write this paper also implies that the prior empirical literature does not provide such benchmarks. As noted in the introduction, most papers in this literature estimate the elasticity of demand for medical care with respect to its price under a specific assumption about how forward looking individuals are. For example, the commonly cited price elasticity of demand for medical care of -0.2 from the RAND Health Insurance Experiment was estimated under the assumption that individual behavior is completely myopic (Keeler and Rolph, 1988), which is precisely the question we are investigating.[23]

Some assumption about the nature and extent of forward looking behavior is required in the existing literature because it has not examined the impact of linear contracts on spending. If we had an estimate of the utilization response to the coinsurance rate in a linear contract, for which the price of medical care is constant for an individual throughout the year, this would be a useful benchmark against which to compare our estimated response to the future price. In a linear contract, dynamic considerations should not affect utilization decisions, so that the behavioral response to different prices (coinsurance rates) would be invariant to the extent of forward looking behavior, and could therefore shed light on the "primitive" substitution between health and income. However, we know of no estimates of the response to a linear contract, nor a source of clean variation in the (constant) coinsurance rate that could be used to identify this response.[24] In the remainder of this section, we therefore calibrate a stylized dynamic model in order to translate our baseline estimate of the response to the future price into economic primitives of interest.

Calibration exercise To try to gauge what our estimated elasticity with respect to the future price implies for how forward looking individuals are, as well as to assess the implications of this finding for the impact of alternative health insurance contracts on *annual* medical spending, we turn to the stylized model of medical utilization decisions in response to health shocks that we

[23] More precisely, Keeler and Rolph (1988) assume that individuals are completely myopic about the possibility of future health shocks in making current medical spending, but that they have perfect foresight regarding all of the year's medical spending associated with a given health shock.

[24] In the Appendix we show how we can use the experimental variation from the RAND Health Insurance Experiment in both the coinsurance rate and the out-of-pocket maximum to extrapolate (out of sample) to the effect of the coinsurance rate in a plan with an infinite out-of-pocket maximum, which thus approximates the response to a linear contract. Our point estimates, while quite imprecise in most specifications, tend to suggest a semi-elasticity of medical utilization with respect to the price of a linear contract that ranges from our estimate of the semi-elasticity with respect to the future price to up to twice as large as this estimate. We interpret the results of this exercise as suggestive of potentially substantial, but perhaps not full, forward looking behavior.

developed in Section 2. We investigate what degree of forward looking behavior (δ) is needed in that model to generate the magnitude of the response of initial medical utilization to the future price that we estimated in our data. Specifically, we calibrate the other parameters of the model and then simulate the response of initial medical utilization to the future price under alternative assumptions about δ; we search for the value of δ that, in this calibrated model, produces the response to the future price that we estimated in the foregoing empirical work.

To do this exercise requires that we calibrate the other primitives of the model in Section 2. These are the arrival rate λ of medical shocks, and the distribution of medical shocks $G(\theta, \omega)$ when they arrive. The latter can be rewritten as $G(\theta, \omega) \equiv G_2(\omega|\theta)G_1(\theta)$. That is, $G_1(\theta)$ represents the unconditional distribution of the total spending that would be required to treat medical shocks and $G_2(\omega|\theta)$ represents the distribution of the (monetized) utility loss from not treating a medical shock of size θ; in that sense, the distribution of ω relative to θ (or simply the distribution of the ratio ω/θ) can be thought of as the "primitive" price elasticity that captures substitution between health and income. As ω/θ is higher (lower), the utility loss is greater (smaller) relative to the cost of treating the shock, so (conditional on the price) the medical shock is more (less) likely to be treated.

We draw on data from the RAND Health Insurance Experiment to calibrate these additional parameters.[25] Conducted over three to five years in the 1970s on a representative population of individuals under 65, the key feature of the RAND experiment was to experimentally vary the health insurance plans to which individuals were assigned. In particular, the coinsurance in the plans varied from "free care" (zero coinsurance rate) to 100% coinsurance rate, with individuals also assigned to 25%, 50%, and 95% coinsurance rates. The details of the experimental design as well as the main results in terms of the impact of consumer cost sharing on healthcare spending and health have been summarized elsewhere (Manning et al., 1987; Newhouse et al., 1993).[26] The estimates from this famous study still remain the standard reference for calibration exercises that require a moral hazard estimate for health insurance (e.g., Finkelstein, Luttmer, and Notowidigdo, 2008; Mahoney, 2010; Gross and Notowidigdo, 2011) and the standard benchmark with which to compare newer estimates of the impact of health insurance on health spending (e.g., Finkelstein, 2007; Chandra, Gruber, and McKnight, 2010; Finkelstein et al., 2011).

Two features of the RAND experiment are very useful for our particular calibration exercise. First, the existence of detailed data on medical claims under a zero cost sharing (free care) plan is not something that, to our knowledge, exists elsewhere. Such data allow us to calibrate the distribution of medical shocks (λ and $G_1(\theta)$) from data that is "uncensored" by any response to cost-sharing; by contrast, any other plan with positive consumer cost sharing only provides information on the medical shocks that are endogenously treated. Second, the experimental variation in plan assignment helps us calibrate the primitive price elasticity $G_2(\omega|\theta)$.

[25] The data from the RAND experiment have, very helpfully, been made publicly available by the RAND investigators through ICPSR.

[26] In the Appendix we provide some more details on the experimental design and the data.

We defer many of the calibration details to the Appendix, and only summarize them here briefly. In the first step of our calibration exercise, we perform a simple statistical exercise to calibrate the weekly arrival and distribution of medical shocks (λ and $G_1(\theta)$) based on the detailed utilization data for the approximately 2,400 family-years we observe in the RAND's free care plan.

In the second step, we use the experimental plan variation in the RAND data to calibrate $G_2(\omega|\theta)$. As mentioned above and discussed in more detail in the Appendix, the RAND experiment does not involve variation in linear contracts that would allow us to directly estimate the "primitive" price elasticity $G_2(\omega|\theta)$. Rather, families in the experiment were randomized into plans with different coinsurance rates and then, within each positive coinsurance rate, they were further randomized into plans with different out-of-pocket maximums. The observed changes in behavior, as both the coinsurance rate and the out-of-pocket maximum are experimentally varied, are therefore influenced by both $G_2(\omega|\theta)$ and δ. Our second step of the calibration exercise uses the random assignment of families to plans and our calibrated model of the arrival and distribution of medical shocks, to map the spending response to different plans to values of $G_2(\omega|\theta)$ and δ that would rationalize this spending response. Fortunately, the resultant values of $G_2(\omega|\theta)$ are quite stable, and are not at all sensitive to the value of δ, so that we can use the RAND experiment to calibrate $G_2(\omega|\theta)$ without knowledge of δ.[27] We can thus use the experimental variation to calibrate $G_2(\omega|\theta)$, and are left with δ as the only remaining unknown primitive.

In the final step of the calibration exercise, we use the calibrated parameters of the model that we have just described to simulate initial medical utilization under deductible contracts with coverage horizons of 3 to 11 months, artificially replicating the setting in which we obtained our estimated elasticity of initial medical utilization with respect to the future price. We repeat this simulation under alternative assumptions about the value of δ. Higher values of δ correspond to greater changes in initial medical utilization as the coverage horizon varies. To quantify this, we regress, for each δ, initial medical utilization in the simulated data on the future price. We then ask what value of δ gives rise to the magnitude of the change in initial medical utilization with respect to the future price that we estimated based on variation in the coverage horizon in our employer-provided data (see last two rows of Table 5).

Calibrated value of δ Figure 2 illustrates our exercise by plotting the semi-elasticity of initial (three month) medical spending with respect to the future price implied by each value of δ. Our

[27] Less fortunately, the converse is not true: the RAND experiment by itself does not allow us to pin down δ with any confidence. In principle, the experimental variation in both coinsurance rates and out of pocket maximums makes the RAND data seem perfectly suited to test and quantify forward looking behavior (since there is experimental variation in the future price conditional on the experimentally determined spot price). In practice, however, using the RAND data to estimate the behavioral response to the future price encounters two important obstacles. The first is conceptual: the combination of non-trivial risks of fairly large expenditure shocks and a preponderance of relatively low out-of-pocket maximums means that is difficult to isolate variation in the future price, as it mechanically generates variation in spot prices that is driven by large medical shocks that are greater than the (lower) out of pocket maximums. The second obstacle is practical: given its much smaller sample size, our attempt to use the RAND variation (despite the first issue) to estimate the behavioral response to the future price produced extremely noisy estimates. The Appendix provides additional details and results of this analysis.

point estimate of the relationship between initial medical spending and future price was -1.08 in the OLS estimation in the penultimate row of Table 5. with the 95% confidence interval ranging from -1.66 to -0.50. Figure 2 indicates that this point estimate in the simulated data is achieved with $\delta = 0.2$. with the 95% confidence interval ranging from 0.06 to 0.45. Table 7 summarizes the implied δ's from the simulation exercises using the alternative dependent variable (time to first claim) and based on IV estimation rather than OLS estimation in both the actual and simulated data.[28] Across the four specifications. the point estimate of δ are centered around 0.2. with a low of around 0.1 and a high of around 0.7.

These calibration results therefore suggest that while we find evidence of forward looking behavior, the extent of forward looking behavior is far from what would be implied by a perfectly rational, fully forward looking individual ($\delta \approx 1$) and closer to what would be implied by a fully myopic individual ($\delta = 0$). Of course, as we noted at the outset, δ – or "forward looking" behavior in our context – should not be interpreted as a pure rate of time preference; liquidity constraints and/or imperfect understanding of the coverage details can push the estimated δ below the rate of time preference, and presumably do so in our context.

Implications for impact of health insurance on spending behavior We can also use our calibrated model to try to assess whether the positive but low δ we have calibrated is quantitatively important for understanding the response of medical utilization to non-linear health insurance contracts. In other words, we try to get a feel for whether, despite the fact that our testing exercise in the main part of the paper rejects fully myopic behavior, myopia could be a reasonable way to approximate behavior. The answer will depend of course not only on our estimate of δ but also on the other parameters of the model and the contracts examined. For example, if the deductible level is low and the vast majority of individuals will exhaust it quickly, most individuals will spend most of the time past the deductible, where they are effectively covered by a linear contract, so that the extent of forward looking behavior would not matter much for the impact of the health insurance contract on medical utilization.

Figure 3 uses the calibrated model to report total annual spending for contracts with different deductible levels in the range of what is common in employer-provided health insurance contracts, and full coverage (zero coinsurance rate) beyond the deductible. It shows results under alternative assumptions about δ. The annual spending levels are based on simulated results from the calibrated model. We are interested in whether low values of δ (of, say, 0.1 or 0.2) can be reasonably approximated by an assumption of either complete myopia ($\delta = 0$) – as underlies for example the famous RAND estimate of the price elasticity of demand for medical care – or an assumption of perfectly forward looking behavior ($\delta \approx 1$) – as has been assumed by other papers estimating the responsiveness of medical care to health insurance contracts. The results in the figure suggest that both these extremes produce substantively different results for the impact of these health insurance

[28] Since the endogeneity of the measured future price to initial medical utilization exists in both the actual and simulated data, comparing the OLS estimates from the actual data to the OLS estimates of the simulated data – or comparing the IV estimates from the actual data to the IV estimates from the simulated data – are both meaningful.

contracts on total spending relative to our calibrated estimates of δ. For example, across all the deductible levels we consider, as we move from the no-deducible plan to a positive deductible plan the decrease in spending implied by $\delta = 0.2$ is 25 to 50 percent smaller than what would be implied by myopic behavior ($\delta = 0$), and 50 to 270 percent greater than what would be implied by $\delta = 1$. These results point to the empirical importance of accounting for dynamic incentives in analyses of the impact of health insurance on medical utilization, and relatedly to the dangers in trying to summarize health insurance contracts with a single price.

2.5 Conclusion

Our paper rejects the null of completely myopic behavior in individuals' response to the non-linear price of medical care. This result jointly indicates that individuals understand something about the non-linear pricing schedule they face, and that they take account of the future price of care in making current medical decisions. Calibration results from our stylized, dynamic model of medical utilization suggest that, at least in the populations we study, individuals may be far from fully forward looking, but that, nonetheless, the extent of forward looking behavior we detect has a non-trivial impact for forecasting how medical spending will respond to changes in non-linear health insurance contracts.

These findings have important implications for estimating or forecasting the impact of alternative health insurance contracts on medical spending, which is a topic that receives great interest and attention both by academics and in the public policy arena. As we noted at the outset, the work to date has almost exclusively focused on estimating (and then using) the elasticity of demand for medical care with respect to its price. However, faced with a non-linear budget set, unless individuals are completely myopic or completely forward looking in their decision making, characterizing moral hazard in health insurance using a single elasticity estimate is neither informative as to how it should be used (relative to which price?) nor is it conceptually well-defined (there are at least two price elasticities that are relevant). Thus, our results highlight the need for more complete modeling of medical utilization induced by the health insurance contract in estimating and forecasting the likely effects of these non-linear pricing schedules among forward looking individuals. More generally, our results speak to the question of whether individuals understand and respond to the incentives embodied in non-linear pricing schedules, of which health insurance contracts are just one of many common examples.

Of course, our findings are specific to our population, which consists of individuals with employer-provided health insurance. Such individuals may be more forward looking than the general population, or may be less liquidity constrained and therefore less responsive to the spot price. It is therefore very possible that in other populations, particularly populations with lower education or income, the extent or even the existence of forward looking behavior might be very different. In settings where individuals appear to behave mostly or entirely myopically it becomes both interesting and important to understand the sources of this apparent myopia, such as the relative roles of time horizon and liquidity constraints. We think that extending our analysis to other settings and attempting to decompose the sources of any myopic component of behavior are promising directions for future work.

References

Adams, William, Liran Einav, and Jonathan Levin (2009). "Liquidity Constraints and Imperfect Information in Subprime Lending." *American Economic Review* 99(1), 49-84.

Blundell, Richard, and Thomas MaCurdy (1999). "Labor Supply: A Review of Alternative Approaches" in Ashenfelter, Orley, and David Card (eds.), *Handbook of Labor Economics.* Oxford: Elsevier North Holland.

Borenstein, Severin (2009). "To What Electricity Price Do Consumers Respond? Residential Demand Elasticity Under Increasing-Block Pricing." Mimeo, UC Berkeley.

Burtless, Gary, and Jerry Hausman (1978). "The Effect of Taxation on Labor Supply: Evaluating The Gary Negative Income Tax Experiment." *Journal of Political Economy* 86(6), 1103–1130.

Cardon, James H., and Igal Hendel (2001). "Asymmetric Information in Health Insurance: Evidence from The National Medical Expenditure Survey." *Rand Journal of Economics* 32, 408-427.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight (2007). "Patient Cost-Sharing, Hospitalization Offsets, and the Design of Optimal Health Insurance for the Elderly." NBER Working Paper No. 12972.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight (2010). "Patient Cost-Sharing, Hospitalization Offsets, and the Design of Optimal Health Insurance for the Elderly." *American Economic Review* 100(1): 193-213.

Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri (2011). "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics* 126(2), 749-804.

Chetty, Raj, and Emmanuel Saez (2009). "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients." Mimeo, Harvard University.

Chevalier, Judith, and Austan Goolsbee (2009). "Are Durable Goods Consumers Forward-Looking? Evidence from College Textbooks." *Quarterly Journal of Economics* 124(4), 1853-1884.

Currie, Janet, and Jonathan Gruber (1996). "Health Insurance Eligibility, Utilization of Medical Care, and Child Health." *Quarterly Journal of Economics* 111(2), 431-466.

Eichner, Matthew J. (1997). "Medical Expenditures and Major Risk Health Insurance." MIT Ph.D. Dissertation, Chapter 1.

Eichner, Matthew J. (1998). "The Demand for Medical Care: What People Pay Does Matter." *American Economic Review Papers and Proceedings* 88(2), 117-121.

Einav, Liran, Amy Finkelstein, Stephen Ryan, Paul Schrimpf, and Mark R. Cullen (2011). "Selection on Moral Hazard in Health Insurance." NBER Working Paper No. 16969.

Ellis. Randall (1986). "Rational Behavior in the Presence of Coverage Ceilings and Deductibles." *RAND Journal of Economics* 17(2). 158-175.

Feldman. Naomi E., and Peter Katuscak (2006). "Should the Average Tax Rate Be Marginalized?" Working Paper No. 304. CERGE-EI.

Finkelstein. Amy (2007). "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *Quarterly Journal of Economics* 122(1). 1-37.

Finkelstein, Amy, Erzo Luttmer and Matthew Notowidigdo (2008). "What Good Is Wealth Without Health? The Effect of Health on the Marginal Utility of Consumption." NBER Working Paper No. 14089.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group (2011). "The Oregon Health Insurance Experiment: Evidence from the First Year." NBER Working Paper No. 17190.

French, Eric, and John B. Jones (2004). "On the Distribution and Dynamics of Health Costs." *Journal of Applied Econometrics* 19(6), 705–721.

Gross, Tal, and Matthew Notowidigdo (2011). "Health Insurance and the Consumer Bankruptcy Decision: Evidence from Expansions of Medicaid." *Journal of Public Economics* 95(7-8), 767-778.

Grubb, Michael D., and Matthew Osborne (2011). "Cellular Service Demand: Tariff Choice, Usage Uncertainty, Biased Beliefs, and Learning." Mimeo, MIT.

Handel, Benjamin (2011). "Adverse Selection and Switching Costs in Health Insurance Markets: When Nudging Hurts." Mimeo, UC Berkeley.

Hausman, Jerry (1985). "The Econometrics of Nonlinear Budget Sets." *Econometrica* 53, 1255–1282.

Ito, Koichiro (2010). "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." Mimeo, UC Berkeley.

Keeler, Emmett, Joseph P. Newhouse, and Charles Phelps (1977). "Deductibles and The Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty." *Econometrica* 45(3), 641-655.

Keeler, Emmett B., and John E. Rolph (1988). "The Demand for Episodes of Treatment in the Health Insurance Experiment." *Journal of Health Economics* 7, 337-367.

Kowalski. Amanda (2010). "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care." NBER Working Paper No. 15085.

Kowalski. Amanda (2011). "Estimating the Tradeoff Between Risk Protection and Moral Hazard with a Nonlinear Budget Set Model of Health Insurance." Mimeo. Yale University.

Liebman. Jeffrey B., and Erzo F. P. Luttmer (2011). "Would People Behave Differently If They Better Understood Social Security? Evidence From a Field Experiment." NBER Working Paper No. 17287.

Liebman. Jeffrey B., and Richard J. Zeckhauser (2004). "Schmeduling." Mimeo. Harvard University.

Manning. Willard. Joseph Newhouse. Naihua Duan. Emmett Keeler. Arleen Leibowitz. and

Susan Marquis (1987). "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review* 77(3), 251-277.

Mahoney, Neale (2010). "Bankruptcy as Implicit Health Insurance." Unpublished mimeo. Available at http://www.stanford.edu/~nmahoney/Research/Mahoney_Bankruptcy.pdf.

Marsh, Christina (2011). "Estimating Health Expenditure Elasticities using Nonlinear Reimbursement." Mimeo, University of Georgia.

Medstat (2006). "MarketScan Commercial Claims and Encounters Database, Description of Deliverables."

Newey, Whitney K. (1987). "Specification Tests for Distributional Assumptions in the Tobit Model." *Journal of Econometrics* 34, 124-145.

Newhouse, Joseph P., and the Insurance Experiment Group (1993). *Free for All? Lessons from the RAND Health Insurance Experiment.* Harvard University Press, Cambridge, MA.

Reiss, Peter C., and Matthew W. White (2005). "Household Electricity Demand, Revisited." *Review of Economic Studies* 72, 853–883.

Saez, Emmanuel (2010). "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2, 180–212.

Van den Berg, Gerard J. (2001). "Duration Models: Specification, Identification and Multiple Durations." In J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics* (First Ed.), Vol. 5, Amsterdam: Elsevier, Chapter 55, 3381–3460.

Yao, Song, Yuxin Chen, Carl F. Mela, and Jeongwen Chiang (2011). "Determining Consumers' Discount Rates with Field Studies." Mimeo, Duke University.

Zweifel, Peter and Willard Manning. (2000). "Moral hazard and consumer incentives in health care." In A.J. Culyer and J.P. Newhouse (eds.), *Handbook of Health Economics* Vol. 1, Amsterdam: Elsevier, Chapter 8, 410-459.

Figure 2.1: Model illustration

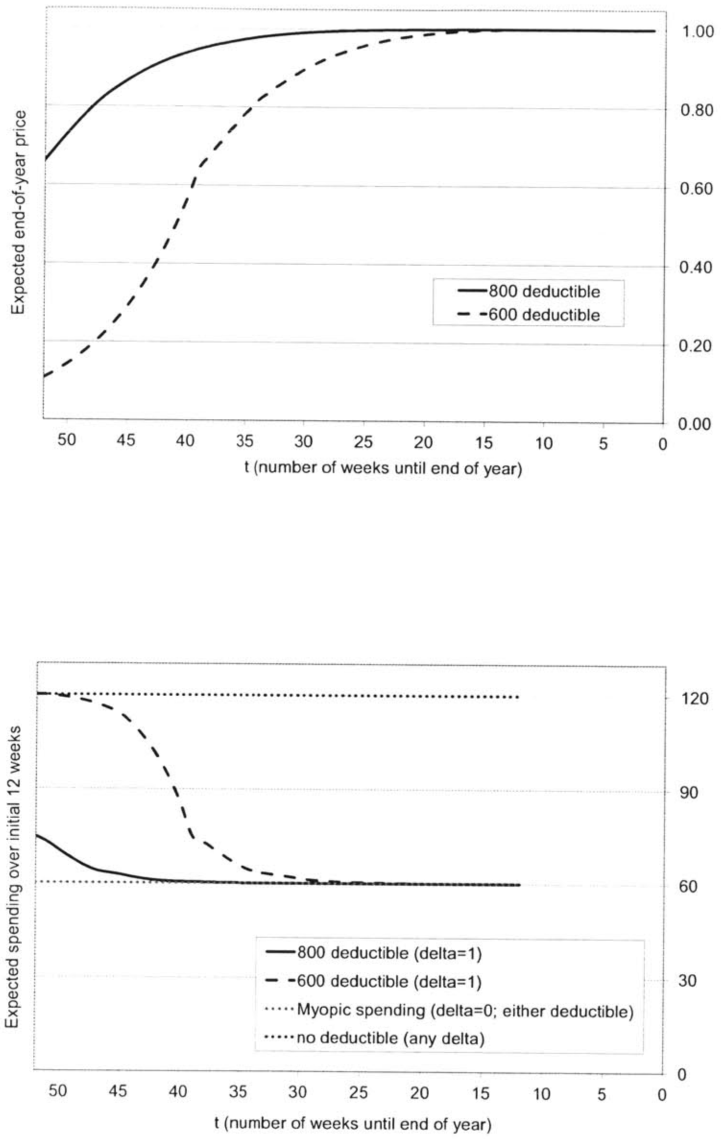


Figure illustrates the implications from a numerical solution to a simple version of the model described in Section 2. We assume $\lambda = 0.2$ and medical events are drawn uniformly from a two-point support of $(\theta = 50, \omega = 50)$ and $(\theta = 50, \omega = 45)$. Expected end-of-year price is equal to one minus the probability of hitting the deductible by the end of the year.

Figure 2.2: Calibration of δ

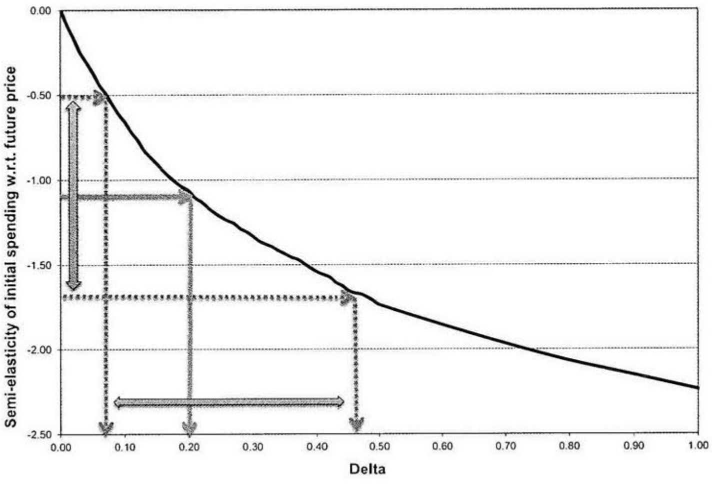


Figure illustrates our calibration exercise. The plot presents the relationship implied by our calibration exercise (see the Appendix for details) between δ and the semi-elasticity of initial medical spending with respect to the future price. The arrows then illustrate how the point estimate and the confidence interval of our semi-elasticity OLS estimate of the impact of the future price on initial spending (penultimate row of Table 5) translate to a point estimate and a confidence interval for δ.

Figure 2.3: The effect of δ on *annual* spending

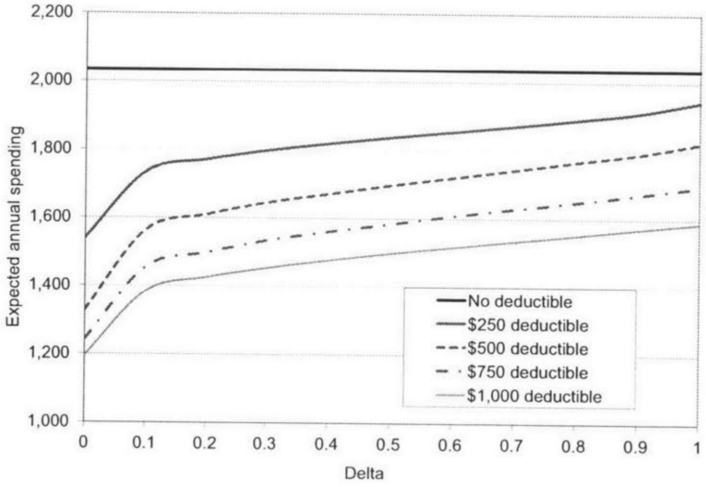


Figure illustrates the implications δ on overall (annual) spending, given the calibration exercise (see the Appendix for details), for a range of possible contracts. The black line represents a case of full insurance, in which overall spending is highest and does not depend on δ. The other lines represent overall spending for deductible contracts which provide full insurance (zero coinsurance rate) once the deductible level has been reached.

Table 2.1: Summary statistics

Employer	Plan	Years Offered	Mid-year new enrollees[a]	In-network deductible ($)		Fraction family	Fraction female	Average age	"Average" enrollment month[b]
				Single	Family				
Alcoa	A0	04-07	3,269	0	0	0.622	0.379	38.56	6.28
	A1	04-07	3,542	250	500	0.408	0.254	35.68	6.42
Firm B	B0	01-05	37,759	0	0	0.530	0.362	36.77	6.35
	B1	01-05	9,553	150	300	0.382	0.341	36.87	6.29
Firm C[c]	C0	99-02, 04-05	27,968	0	0	0.348	0.623	36.40	7.35
	C1	99-00	6,243	200	500	0.348	0.622	37.53	7.50
	C2	01-02	8,055	250	625	0.323	0.606	38.66	7.56
	C3	04-05	5,633	300	750	0.299	0.660	38.51	7.67

[a] The sample includes employees who enroll in February through October.

[b] In computing the "average" enrollment month we number the join months from 2 (February) through 10 (October).

[c] We omit 2003 from the analysis since the plan documentation regarding the deductible plan was incomplete in that year.

Table 2.2: Variation in expected end-of-year price

		Deductible	Expected end-of-year price[a]		
Employer	Plan	(Single/Family)		Joined plan in:	
		[N = enrollees]	Feb-Apr	May-Jul	Aug-Oct
Alcoa	A0	0/0	0.000	0.000	0.000
		[N = 3,269]	(N = 1,007)	(N = 981)	(N = 1,281)
	A1	250/500	0.512	0.603	0.775
		[N = 3,542]	(N = 975)	(N = 1,114)	(N = 1,453)
Firm B	B0	0/0	0.000	0.000	0.000
		[N = 37,759]	(N = 8,863)	(N = 15,102)	(N = 13,794)
	B1	150/300	0.529	0.630	0.806
		[N = 9,553]	(N = 2,165)	(N = 4,175)	(N = 3,213)
Firm C	C0	0/0	0.000	0.000	0.000
		[N = 27,968]	(N = 6,504)	(N = 6,158)	(N = 15,306)
	C1-C3[b]	200-300/500-750	0.543	0.633	0.811
		[N = 19,931]	(N = 4,001)	(N = 4,143)	(N = 11,787)

[a] Expected end-of-year price is equal to the fraction of individuals who do not hit the deductible by the end of the calendar year (and therefore face a marginal price of 1). It is computed based on the plan's deductible level(s), the join month, and the annual spending of all the employees who joined that plan in that month; we compute it separately for family and single coverage within a plan and report the enrollment-weighted average.
[b] In firm C, we pool the three different deductible plans (C1, C2, and C3) that are offered in different years.

Table 2.3: Initial medical utilization by join quarter

Employer	Plan	Deductible (Single/Family) [N = enrollees]	Average initial (first 3 months) spending ($) Joined plan in: Feb-Apr	 May-Jul	 Aug-Oct	Average days to first claim (censored^{a}) Joined plan in: Feb-Apr	 May-Jul	 Aug-Oct
Alcoa	A0	0/0	1,092	1,409	1,270	42.8	43.2	43.7
		[N = 3,269]	(s.d. = 2,679)	(s.d. = 9,217)	(s.d. = 4,733)	(s.d. = 33.2)	(s.d. = 33.4)	(s.d. = 33.06)
	A1	250/500	727	788	451	63.9	63.4	66.7
		[N = 3,542]	(s.d. = 3,730)	(s.d. = 4,324)	(s.d. = 2,216)	(s.d. = 33.4)	(s.d. = 33.6)	(s.d. = 32.7)
Firm B	B0	0/0	628	596	647	57.0	58.3	58.8
		[N = 37,759]	(s.d. = 4,841)	(s.d. = 3,632)	(s.d. = 2,886)	(s.d. = 33.3)	(s.d. = 33.7)	(s.d. = 33.1)
	B1	150/300	723	682	521	65.0	65.5	71.1
		[N = 9,553]	(s.d. = 4,587)	(s.d. = 4,046)	(s.d. = 2,896)	(s.d. = 32.3)	(s.d. = 31.8)	(s.d. = 29.7)
Firm C	C0	0/0	539	546	505	57.1	58.1	57.0
		[N = 27,968]	(s.d. = 3,087)	(s.d. = 2,305)	(s.d. = 3,017)	(s.d. 34.6)	(s.d. = 34.0)	(s.d. = 34.5)
	C1-C3	200-300/500-750	515	581	495	56.1	56.2	57.6
		[N = 19,931]	(s.d. = 1,842)	(s.d. = 2,126)	(s.d. = 2,556)	(s.d. = 35.0)	(s.d. = 34.6)	(s.d. = 34.8)

All utilization measures refer to utilization by the employee and any covered dependents.

[a] Days to first claim is censored for all employees at 92 days. 42% of the observations are censored.

Table 2.4: The relationship between join month and initial medical utilization

		Deductible	Log Initial Spending[a]		Log Time to First Claim[b]	
Employer	Plan	(Single/Family)	Difference	DD	Difference	DD
		[N = enrollees]	(1)	(2)	(3)	(4)
Alcoa	A0	0/0	-0.003		0.007	
		[N = 3,269]	(0.023)		(0.010)	
	A1	250/500	-0.015	-0.012	0.003	-0.005
		[N = 3,542]	(0.021)	(0.027)	(0.014)	(0.015)
Firm B	B0	0/0	-0.015		0.024	
		[N = 37,759]	(0.007)		(0.008)	
	B1	150/300	-0.091	-0.075	0.059	0.033
		[N = 9,553]	(0.026)	(0.025)	(0.014)	(0.010)
Firm C	C0	0/0	-0.004		0.003	
		[N = 27,968]	(0.013)		(0.006)	
	C1-C3	200-300/500-750	-0.027	-0.022	0.019	0.016
		[N = 19,931]	(0.012)	(0.010)	(0.007)	(0.006)

Table reports coefficients (and standard errors in parentheses) from regressing a measure of initial medical care utilization (defined in the top row) on join month (which ranges from 2 (February) to 10 (October)). Columns (1) and (3) report the coefficient on join month separately for each plan, based on estimating equation 4; the regressions also include an indicator variable for coverage tier (single vs. family). Columns (2) and (4) report the difference-in-differences coefficient on the interaction of join month and having a deductible plan, separately for each firm, based on estimating equation 5; the regressions also include plan by coverage tier fixed effects and join month fixed effects. Standard errors are clustered on join month by coverage tier.

[a] Dependent variable is $log(s+1)$ where s is the total medical spending of the employee and any covered family members in their first three months in the plan.

[b] Dependent variable is $log(time)$ where "time"is the number of days to first claim by any covered family member, censored at 92. We estimate the regressions in columns (3) and (4) by Tobit.

Table 2.5: The relationship between expected end-of-year price and initial medical utilization

Sample	N	Log Initial Spending[a]		Log Time to First Claim[b]	
		Difference (1)	DD (2)	Difference (1)	DD (2)
Alcoa	6,811	-0.92 (0.30)	-0.76 (0.51)	0.294 (0.176)	0.046 (0.191)
Firm B	47,312	-2.02 (0.57)	-1.73 (0.54)	1.171 (0.363)	0.677 (0.227)
Firm C	47,899	-0.89 (0.39)	-0.81 (0.37)	0.357 (0.234)	0.254 (0.143)
Pooled (OLS)	102,022		-1.08 (0.29)		0.357 (0.113)
Pooled (IV)	102,022		-0.78 (0.27)		0.564 (0.135)

Table reports coefficients (and standard errors in parentheses) from regressing a measure of initial medical care utilization (defined in the top row) on the expected end-of-year price, computed for each plan (by coverage tier) and join month. Columns (1) and (3) report the coefficient on expected end-of-year price fp separately for each deductible plan in each firm, based on estimating equation 6; the regressions also include an indicator variable for coverage tier (single vs. family). Columns (2) and (4) report the coefficient on expected end-of-year price fp from estimating equation 7, which now includes the no-deductible plans as well; these regressions also include plan by coverage tier fixed effects and join month fixed effects. In the bottom two rows, we report the coefficient on expected end-of-year price fp from estimating equation 8 using OLS and IV (respectively) by pooling the data from all firms and plans; in addition to plan by coverage tier and join month fixed effects, these regressions now also include firm by join month fixed effects. Standard errors are clustered on join month by coverage tier by firm. The IV specification makes use of a simulated end-of-year price as an instrument for the expected end-of-year price (see text for details). The coefficient on the instrument in the first stage is 0.56 (standard error 0.024); the F-statistic on the instrument is 524.

[a] Dependent variable is $log(s+1)$ where s is the total medical spending of the employee and any covered family members in their first three months in the plan

[b] Dependent variable is $log(time)$ where "time" is the number of days to first claim by any covered family member, censored at 92 days. We estimate the regressions in columns (3) and (4) by Tobit.

Table 2.6: Robustness and specification checks

Specification		N	Initial Spending		Time to First Claim	
			Coeff on fp	Std. err.	Coeff on fp	Std. err.
			(1)	(2)	(3)	(4)
(1)	Baseline (logs)	102,022	-1.08	(0.29)	0.357	(0.113)
(2)	Level	102,022	-394.43	(162.12)	14.842	(4.429)
(3)	QMLE Poisson	102,022	-0.70	(0.25)	–	–
(4)	OLS (No Tobit)	102,022	–	–	0.114	(0.057)
(5)	Varying censor points	102,022	–	–	0.330	(0.114)
(6)	Cox proportional hazard model	102,022	–	–	-0.347	(0.109)
(7)	Control for demographics	102,014	-0.98	(0.26)	0.524	(0.121)
(8)	Only those who remain for 2nd year	64,398	-0.73	(0.34)	0.161	(0.133)
(9)	Dep. var measured in 2nd year	64,398	-0.17	(0.31)	–	–
(10)	Dep. var measured Jan-Mar of 2nd year	64,398	-0.44	(0.26)	0.172	(0.106)

Table reports results from alternative analyses of the relationship between initial medical utilization and expected end-of-year price. Row (1) shows the baseline results from estimating equation 8 by OLS in columns 1 and 2 and by Tobit in columns 3 and 4, as in the penultimate row of Table 5. Alternative rows report single deviations from this baseline as explained below. In row (2) the dependent variables are defined in levels rather than logs. Mean dependent variables are 596.2 dollars (initial spending) and 58.3 days ((censored) time to first claim). In row (3) the dependent variable is defined in levels (not logs) and the regression is estimated by quasi-maximum likelihood Poisson instead of OLS. In row (4) the regression is estimated by OLS rather than Tobit. In row (5) we estimate the same regression as in the baseline, but we now allow the censoring point to vary with join month, from 92 days if the employee joined in October to 334 days if the employee joined in February. In row (6) we estimate a Cox semi-parametric proportional hazard model on the time to first claim (censored at 92 days); note that here a longer time to first claim is indicated by a negative coefficient (a lower "failure" rate). In row (7) we add controls for age, gender, and start year (as well as interactions of each of those with the firm fixed effects) to the baseline specification. In row (8) we estimate the baseline specification on a smaller sample of employees who remain in the firm for the entire subsequent year; in row (9) we estimate the baseline specification on this same sample, but defining the dependent variable based on utilization in the same three months of the subsequent year (i.e., their first full year in the firm); in row (10) we estimate the baseline specification on this same sample but now define the dependent variable based on utilization in January to March of the first full year in the firm.

Table 2.7: Calibrating δ

	Log(Three Month Spending)		Log(Time To First Claim)	
	OLS	IV	Tobit	Tobit IV[a]
Estimated semi-elasticity				
Point Estimate	-1.08	-0.78	0.36	0.56
CI Lower Bound	-0.50	-0.24	0.14	0.30
CI Upper Bound	-1.67	-1.33	0.58	0.83
Implied delta				
Point Estimate	0.20	0.12	0.26	0.67
CI Lower Bound	0.06	0.02	0.06	0.19
CI Upper Bound	0.45	0.28	0.76	1.00

Panel A reports the estimated semi-elasticities of initial medical utilization with respect to the future price; these are taken directly from the last two rows of Table 5. Panel B shows the implied values of δ associated with each estimate based on the calibration exercise described in the text.

[a] We impose 1 for the upper bound of the confidence interval for the implied δ in the Tobit IV case based on our a priori knowledge that δ cannot be higher than 1; no δ less than 1 produces a semi-elasticity as large as 0.83 in our model.

Appendix

This appendix describes in more detail the uses we make of data from the RAND Health Insurance Experiment.[29] Appendix A describes our attempt to use the RAND experiment random assignment of out-of-pocket maximums as the basis for an additional test for forward looking behavior. Appendix B discusses our attempt to use the RAND data to approximately the "primitive" price elasticity of demand, to serve as a benchmark for our estimated response to the future price. Appendix C provides a detailed explanation of how we use the RAND data for the calibration exercises described in Section 4.

As explained in the main text, the RAND experiment, conducted in 1974-1981, randomly assigned participating families to health insurance plans with different levels of cost sharing. Each plan was characterized by two parameters: the coinsurance rate (the share of initial expenditures paid by the enrollee) and the out-of-pocket maximum, referred to as the "Maximum Dollar Expenditure" (MDE). Families were assigned to plans with coinsurance rates ranging from 0% ("free care") to 100%. Within each coinsurance rate, families were randomly assigned to plans with MDEs set equal to 5%, 10%, or 15% of family income, up to a maximum of $750 or $1,000.[30]

A. Testing forward-looking behavior using the RAND data

The latter feature of the RAND plan assignment process – random assignment of MDEs – would seem to provide an ideal experimental setting for a test of forward looking behavior since it potentially provides random variation in the future price among individuals assigned to the same coinsurance (and hence the same spot price). While differences in MDEs across individual families were due in part to differences in family income, differences in average MDE and average end-of-year price across plans can be treated as randomly assigned. Appendix Table A6 provides sample counts and various summary statistics for the RAND plans.[31] As the table shows, average MDEs were considerably higher in plans where the MDE was set equal to 10% or 15% of family income than in plans where the MDE was set to 5% of income. These differences generated corresponding differences in the share of families hitting the MDE and in expected end-of-year price (columns (5) and (6)).

Columns (8) and (9) of Appendix Table A6 present results from a regression of time to first claim on expected end-of-year price in the RAND. Specifically, we run the regression

$$\text{Log(Time-to-First-Claim)}_f = \beta \cdot fp_j + \gamma \cdot coins_j + X_f'\chi + \epsilon_f, \tag{2.9}$$

where fp is the future price (or expected end-of-year price), *coins* is the coinsurance rate, f indexes

[29] The original RAND investigators have very helpfully made their data publicly available. We accessed the data through the Inter-University Consortium for Political and Social Research.

[30] For a detailed description of the plans and other aspects of the experiment, see Newhouse et al. (1993).

[31] Appendix Table A6 omits the RAND's "individual deductible plans," which had coinsurance rates of 0% for inpatient services and 100% or 95% for outpatient services, because there was no MDE variation within this coinsurance rate structure.

families. j indexes plans. and X_f is a vector of dummy variables for site and start month in the experiment by year.[32] As shown. we run the regression separately for each coinsurance rate group and then pool all groups (or all groups except the free care plan) to maximize power (we also run a specification with a full set of coinsurance rate dummies in place of the coinsurance rate term). We run both OLS and Tobit regressions. where the latter account for censoring of time to first claim at 367 days.

There are two important limitations to this analysis, so that despite its apparent advantages. the RAND variation is in fact inferior to the variation generated by employee hire dates (the primary variation used in the paper). First, as a practical matter, the RAND setting gives us much less power to detect differences in spending by expected end-of-year price. The samples are smaller (with a total sample size of 5,653 family-year observations across all plans, relative to more than 100,000 in the combined employer-provided sample), and there is much less variation in end-of-year price. As a result, our estimates based on the RAND data are quite imprecise. Even in our most inclusive specification (bottom row of Table A6), we can neither reject the null of no response to the future price nor reject a coefficient on the future price of 3, far larger than what we find in the employer-provided data.

Second, conceptually, the variation in the MDE in the RAND data is not as clean for testing for forward looking behavior as the variation in the coverage horizon that we use in the paper. To see this, note that differences in expected end-of-year price are correlated with differences in spot price even under the null hypothesis of no forward-looking behavior. Even if people are fully myopic, families in low MDE plans will meet their MDEs sooner and will spend more of the year facing a 0% spot price. As a result, they will have higher spending even if they are not at all forward-looking.[33] Because 12% of families in the lowest MDE, highest coinsurance rate plan hit the MDE within the very first month of the experiment, this is a concern even when looking just at initial (e.g., one month) spending.

We attempted to solve this problem by using time to first claim as the outcome variable. Unfortunately, however, some of RAND's MDE levels are quite low, so they can affect even the spot prices families face when making decisions about their very first health expenditure. To see this, consider two families in plans with a 100% coinsurance rate. The first family has an MDE of $150, the second an MDE of $300. Suppose that, before either family has any other health expenditures, each experiences a health shock that would cost $300 to treat. The out-of-pocket cost of treating this shock would be $150 for the low MDE family but $300 for the high MDE family, meaning that the low MDE family faces a spot price of only 50% for the episode, compared to 100% for the high MDE family. Hence, the low MDE family will be more likely to treat the

[32] Plan assignment was random only conditional on which of the experiment's six sites a family lived at and when the family enrolled in the experiment. For details, see Newhouse et al. (1993, Appendix B).

[33] In contrast, this is not a problem when using the variation in end-of-year price generated by month of hire. If people are fully myopic, then early hires will have the same levels of three-month spending as late hires, and so the two groups will be equally likely to hit their deductibles within three months and will face the same average spot price.

episode. even if both families are fully myopic.

Because about half of outpatient episodes (defined as in Appendix B below) are larger than the smallest MDEs in the RAND sample. this problem is potentially quite significant. Indeed. in simulations mimicking the RAND setting. we obtain a large and statistically significant coefficient on end-of-year price in a regression for time to first claim. even when we assume complete myopia. Thus. we conclude that. even apart from the precision problems. we cannot use the RAND setting to generate variation in the future price conditional on the spot price to test for forward looking behavior.[34]

B. Approximating the "primitive" price elasticity using RAND data

Before turning to the model-based calibration exercise in the next sub-section, we first present a loose way of trying to gauge the extent of forward looking behavior by using the experimental variation in contracts in the RAND data to generate an estimate of the "primitive" price elasticity of medical care utilization which we then compare to our previously estimated response to the future price from the main empirical work in the paper.

The variation used in the main empirical work in the paper is not useful in this regard, as we observe neither linear contracts nor identifying variation for plan assignment. The RAND experiment does not provide this ideal variation either, since all of the RAND contracts (except for the free care contract) involve a non-linear pricing schedule; families were randomized into coinsurance rates and then, within each positive coinsurance rate, they were further randomized into plans with an out-of-pocket maximum (known as the "maximum dollar amount" or MDE in the RAND context) of either 5%, 10%, or 15% of income (up to a maximum of \$1,000 or \$750); above the MDE the price of care is zero.[35] However, RAND's experimental variation (within each coinsurance rate) in the out-of-pocket maximum allows us to estimate its effect, and then to extrapolate out of sample to obtain the behavioral response to a contract where the out-of-pocket maximum is sufficiently high, thus approximating a linear contract.

Specifically, we estimate the regression

$$y_{fj} = \eta_1 \cdot coins_j + \eta_2 \cdot Share\_Hit_j + \eta_3 \cdot coins_j \cdot Share\_Hit_j + v_{fj}, \tag{2.10}$$

where y_{fj} is a measure of medical utilization by family f in plan j, $coins_j$ is the coinsurance rate of the plan the family was randomized into (which is either 0%. 25%. 50%. or 95%). and $Share\_Hit_j$

[34] Two of the original RAND investigators, Keeler and Rolph (1988), also attempt to use the RAND data to test for forward looking behavior, but they use a different empirical strategy. They do not exploit the MDE variation, and instead rely on within-year variation in how close families are to their MDEs. They test whether spending is higher among families who are closer to hitting their MDEs, as would be expected - all else equal - if people are forward looking. They make several modeling assumptions to try to address the (selection) problem that families with higher underlying propensities to spend are more likely to come close to hitting their MDEs. They also assume that individuals have perfect foresight regarding all the subsequent medical expenses within a year associated with a given health shock. They conclude that they cannot reject the null of complete myopia with respect to future health shocks.

[35] All dollar amounts are reported in current (1970s) dollars.

is the fraction of families within the same coinsurance rate and MDE assignment that hit (i.e., spent past) the MDE during the year. For the positive coinsurance plans this number ranges from 8 percent to 40 percent depending on the plan assignment (see Appendix Table A6, column (5)). The coefficient of interest is η_1, which we interpret as the responsiveness of medical utilization to a change in the coinsurance rate of a linear contract; this involves extrapolating out of sample to where $Share\_Hit_j = 0$, which would be the case for a sufficiently high MDE.

Because the share of families in a given plan assignment that hit the MDE depends on family spending behavior, which itself may be endogenous to plan assignment, we also present IV specifications in which we instrument for the share of families in a given plan that hit the MDE with the simulated share hitting the MDE. The "simulated share" is calculated as the share of the entire (common) sample of individuals across all plans that would have hit the MDE if assigned to the given plan, in a similar spirit to the IV exercise we reported in the previous section.

Appendix Table A7 presents the results. Our sample size is approximately 1,500 families (about 5,600 family-years).[36] As in the previous analysis, we analyze both the responsiveness of the first three months of spending and the time to first claim. Here, we also add total (annual) spending as an additional outcome (as the proportional response to a linear contract should not, in principle, be different for initial and total spending).

The response to the linear coinsurance – while fairly imprecise in most specifications – can now be compared to our estimates of the response to the future price from the previous sub-section. Using the IV specification, we find a spending semi-elasticity with respect to the price of a linear contract that ranges from -1.2 to -1.7, which is roughly twice as large as the semi-elasticity of -0.78 with respect to the future price that we found in last section (see last row of Table 5). Similarly, we estimate that the semi-elasticity of the time to first claim with respect to the price of a linear contract is 0.53, which is virtually identical to our analogous semi-elasticity estimate of 0.56 with respect to the future price. Thus, overall the results are indicative of substantial, but perhaps not full, forward looking behavior.

C. Model calibration

In Section 4 of the main text, we explain how we use the RAND data to calibrate a model that allows us to map the estimated elasticity of initial spending with respect to the end-of-year price to the parameter δ. Here, we provide more details about this calibration exercise.

Calibrating the medical shock process (λ and $G_1(\theta)$) We calibrate the medical shock process using data from the 620 families (approximately 2,400 family-years) participating in the RAND's "free care" plan. We calibrate the distribution of inpatient and outpatient shocks separately and

[36] Appendix Table A6 shows the exact plans we study and the distribution of families across those plans. The entire RAND experiment involved about 2,400 families. We exclude from this analysis the approximately 400 families randomized into the 95% coinsurance plan with a fixed ($150 per person) MDE plan (also know as the "individual deductible" plan) because for this MDE only the coinsurance rate differed (it was 95% for outpatient care but free for inpatient care), and the approximately 400 families randomized into an HMO.

also allow for heterogeneity across families in the distribution of shocks. Specifically, letting f index families and ζ index types of spending (inpatient or outpatient), we assume that in each period t, family f draws a shock of type ζ with probability $\lambda_{f\zeta}$. In periods where a family does experience shocks of type ζ, the shocks are drawn i.i.d. from a lognormal distribution with mean $\mu_{f\zeta}$ and variance σ.

Our procedure for obtaining the various spending distribution parameters is as follows. We define a period t as a week. We group together all claims of a given type separated by less than one week and define each grouped set of claims as one episode, assigning it to the first week of the episode; this generated about 6,000 inpatient episodes and about 77,000 outpatient episodes over the course of the entire experiment. For each family and each type of spending, we then compute: $\lambda_{f\zeta}$ as the share of weeks (over the course of the entire experiment[37]) in which family f experienced an episode of type ζ, we set $\mu_{f\zeta}$ as the average size of family f's episodes of type ζ, and $\sigma_{f\zeta}$ as the variance of family f's episodes of type ζ. Because $\sigma_{f\zeta}$ is extremely noisy (even more so than $\mu_{f\zeta}$) and because it is unavailable for families with only one shock of a given type, we set σ to be the average of $\sigma_{f\zeta}$ for all families.

Partly to reduce noise and partly to make simulating the model computationally feasible, we next divide families into five-percentile groups based on their values of λ. We replace each value of $\lambda_{f\zeta}$ and $\mu_{f\zeta}$ with the mean of the respective variable for family f's percentile group. This approach eliminates cases where the probability of outpatient shock is zero, but leaves 55% of the sample with a zero probability of inpatient medical shocks. This is consistent with our intuition that every family faces some meaningful risk of experiencing an outpatient shock, but some families (specifically, those who experience no inpatient episodes at any point during the experiment) may face so little risk of an inpatient episode that they perceive it as approximately zero.

Appendix Figure A1 and Appendix Table A8 compare the actual distributions of expenditures in the free care plan with the simulated distributions. Appendix Figure A1 presents a histogram of total (the sum of inpatient and outpatient) spending, while Appendix Table A8 reports the means and standard deviations of log inpatient, outpatient, and total spending, as well as the share of families with no inpatient, outpatient, or total spending over the course of a year ($T = 52$).[38]

The fit is notably better for outpatient than inpatient spending, basically because, as others have also found (see, e.g., French and Jones, 2004), the lognormal distribution is a better fit for outpatient than inpatient spending. Nonetheless, the fit is fairly good for both categories of spending, and seems (to us) to capture the main properties of health spending for the purpose of our calibration exercise.

Calibrating the distribution of valuations ($G_2(\omega/\theta)$) Recall that ω represents the (monetized) health cost of a given shock, and so ω/θ represents the health cost of a given shock relative to the cost of treatment. For example, if $\omega/\theta = 0.5$ then the health cost of not treating a given

[37] Families participated in the experiment for periods of either three or five years.

[38] Throughout, we define log spending as $log(spending + 1)$ in order to avoid missing values.

shock is equal to half the cost of the treatment.

We calibrate the distribution of ω/θ for outpatient shocks, but assume $\omega/\theta = 1$ for all inpatient shocks. That is, we assume that individuals treat all inpatient shocks, regardless of what share of the cost of treatment they pay out of pocket. This analytic choice is done primarily to make the calibration exercise much more feasible (inpatient shocks are sufficiently rare relative to outpatient shocks that it is much harder to use the data to calibrate ω/θ for them). It also reflects our intuition that most health shocks for which treatment requires hospitalization are much less discretionary than outpatient care; this is consistent with the basic findings from the RAND experiment itself (Newhouse et al., 1993) as well as subsequent quasi-experimental evidence (e.g., Einav et al., 2011) and our findings in this paper that only outpatient care appears to respond to the future price (Appendix Table A2).

For outpatient shocks, we assume that ω/θ follows a Beta distribution with parameters a and b, so that $\omega/\theta \sim \beta(a, b)$. Thus, $a > 0$ and $b > 0$ are the key primitive price elasticity parameters of the model. The ratio $a/(a+b)$ gives the mean value of ω/θ. We use data on the 95%, 50%, and 25% coinsurance RAND plans to calibrate a and b.[39] We simulate the model described in Section 2 to generate utilization data for each coinsurance rate. We then try to match three moments of the actual RAND data for each coinsurance rate: the mean of log spending, the standard deviation of log spending, and the share of the sample with zero spending.[40] Specifically, we minimize the sum of squared differences, weighting by the different coinsurance rates' RAND sample sizes.

So far, we have glossed over a tension with our calibration strategy. Namely, that the distribution of spending (using the model of Section 2) depends not only on the distributions of λ, θ, and ω/θ, but also on δ. And yet our goal is to obtain the parameters of the ω/θ distribution without knowing δ so that we can then determine what value of δ yields the elasticities we obtained from the employer-provided data.

Our strategy succeeds simply because it happens that the objective function is quite flat in δ but quite steep (and generally invariant to δ) in a and b. Panel A of Appendix Table A9 shows, for 11 values of δ ranging from zero to one, the optimal values of a and b and the resulting values of the objective function. As the table shows, the model selects very similar values of a and b regardless of the assumed value of δ, and yields similar values of the objective function (at the optimal values of a and b). Basically, whatever the choice of δ, the best fit involves $E[\omega/\theta] \approx 0.55$ and a highly bimodal distribution for ω/θ, with modes near 0 and 1.[41]

Based on eyeballing the simulation results, we select $a = 0.3$ and $b = 0.25$ for our calibration exercise; these are the values that minimize the objective function averaged over the possible values of δ we examine. Panel B of Appendix Table A9 shows that these values yield a fairly tight fit to

[39] We do not make use of the data from the "mixed coinsurance rate" plans included in Appendix Table A6. Incorporating these plans would have required further complicating the model in order to introduce multiple types of outpatient spending.

[40] As before, we define log spending as $log(spending + 1)$ to avoid missing values.

[41] Intuitively, the bimodal distribution reflects the fact that the sample means from the RAND data are almost the same for the 25% and 50% coinsurance rate plans, implying that, for most outpatient shocks, people either will not treat the shock at a coinsurance rate of 25% or will treat it unless the coinsurance rate is quite high.

the RAND data for any assumed value of δ.

Mapping the elasticity of initial spending with respect to end-of-year price to δ Having calibrated the key elements of the model. the final step in our calibration exercise is to simulate the data generating process from our employer-provided data and obtain estimates of the responsiveness of initial spending to the expected end-of-year price in the simulated data.

We consider plans with deductible of \$0. \$250. \$750. and \$1.000. with no cost-sharing above the deductible. For each of these plans. we use the calibrated parameters described above. and a range of values for δ (the only remaining free parameter). and simulate spending given time horizons of 47. 42. 37. 32. 27. 22. 17. or 12 weeks (analogous to hire dates ranging from February to October). For each of 10,000 simulated families in each deductible-horizon combination. we obtain simulated spending in the first 12 periods (analogous to first three month spending) and time to first claim (here measured in weeks and censored at 12); in addition. for each deductible/horizon combination. we obtain average "end-of-year"price (here. just average price at the end of the horizon).

Letting d denote levels of the deductible. h index horizon lengths. and f index families. we use the simulated data to estimate the regression

$$Outcome_i = \beta \cdot fp_{dh} + \gamma_d + \epsilon_f. \tag{2.11}$$

Here. β is the coefficient of interest, while the γ_d's are dummy variables for deductible level. We estimate the regression for log("three month") spending (spending in the first 12 periods) and for time to first claim. For reasons explained in the main text. we run both OLS and IV regressions. in the latter case instrumenting for fp_{dh} with the average end-of-period price after h periods among families with the maximum time horizon (47 weeks).

We repeat the above exercise for 101 values of δ ranging from 0 to 1. We can then obtain point estimates and confidence intervals for δ by comparing the estimates of the responsiveness of initial spending to end-of-year price obtained in the simulations with the estimates and the bounds of the confidence intervals obtained from the employer-provided data.[42] The results are presented in Figure 2 in the main text.

[42] Technically, the confidence intervals on δ should also take into account the standard errors on β from the regressions in the simulated data. However, because we can make the simulations so large – we simulate 10,000 families for each deductible horizon paid – the standard errors on β are effectively zero. β simply describes the relationships imposed by the model and the calibrated parameters.

Figure A1: Fit of the calibration exercise of medical events

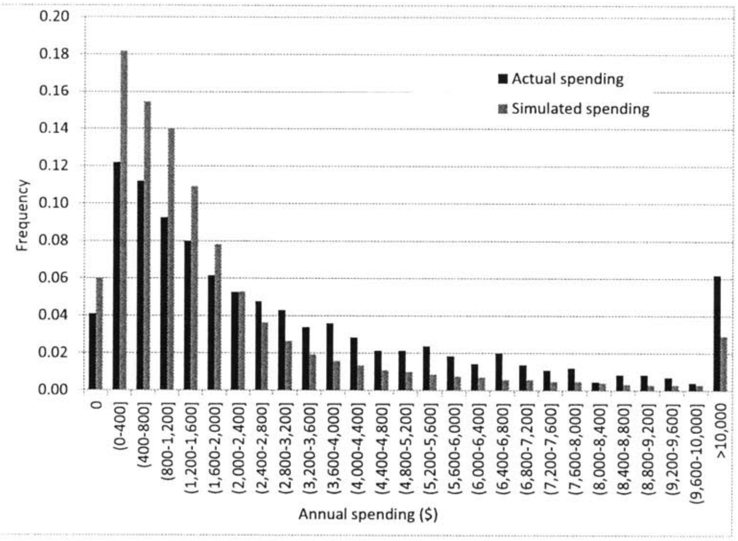


Figure shows the distribution of annual medical spending in the "free care" RAND data based on the actual (black bars) and simulated (gray bars) data. The simulations use the calibrated parameters, as explained in the Appendix. The actual data is based on the 2,376 family-years of data in the free care plan.

Appendix Table A1: Additional plan details

Employer	Plan	Years	Mid-year new enrollees[a]	In-network features						Out-of-network features					
				Deductible ($)		Coins[b]	Copay ($)	Stop loss ($)		Deductible ($)		Coins[b]	Copay ($)	Stop loss ($)	
				Single	Family			Single	Family	Single	Family			Single	Family
Alcoa	A0	2004-07	3,269	0	0	0.10	0	2,500	5,000	250	500	0.3	0	5,000	10,000
	A1	2004-07	3,542	250	500	0.10	0	2,750	5,500	500	1,000	0.3	0	5,500	11,000
Firm B	B0	2001-05	37,759	0	0	0.00	15	0	0	250	500	0.2	0	1,250	2,500
	B1	2001-05	9,553	150	300	??	??	??	1,100	??	??	??	0	??	??
Firm C	C0	1999-05	27,968	0	0	0.00	15	0	0	300	750	0.3	0	3,000	6,000
	C1	1999-00	6,243	200	500	0.10	0	1,000	2,000	??	??	0.3	0	3,750	7,500
	C2	2001-02	8,055	250	625	0.10	0	1,250	2,500	250	625	0.3	0	3,900	7,800
	C3	2004-05	5,633	300	750	0.10	0	1,300	2,600	300	750	0.3	0	3,900	7,800

"??" denotes an unknown feature of a plan.

[a] The sample includes employees who enroll in February through October.

[b] Coinsurance denotes the fraction of medical expenditures the insured must pay out of pocket after hitting the deductible and prior to reaching the "stop loss."

Appendix Table A2: Responsiveness of different types of care to the future price

Dependent variable	Mean of the dep. var.	Coeff on future price	Std. Error
(1) Log initial spending	3.32	-1.08	(0.29)
(2) Log initial outpatient spending	3.29	-1.06	(0.29)
(3) Initial spending	596.2	-394.4	(162.1)
(4) Initial outpatient spending	445.0	-375.8	(107.7)
(5) Initial inpatient spending	147.5	-19.8	(99.1)
(6) Any initial inpatient spending	0.014	-0.008	(0.006)

Table reports results for different types of medical spending of the analysis of the relationship between initial medical spending and expected end-of-year price ("future price"). All rows show the results from estimating equation 8 by OLS using different dependent variables; in addition to "future price" the covariates in this regression include plan by coverage tier fixed effects, join month fixed effects and firm by join month fixed effects. Standard errors are clustered on join month by coverage tier by firm. The first row shows the baseline results (see penultimate row in Table 5) for the dependent variable log initial spending (plus 1). In row 2 the dependent variable is the log of initial outpatient spending (plus 1). Rows 3 through 5 show results for the level of initial medical spending, the level of initial outpatient spending and the level of initial inpatient spending respectively. The last row shows the results for an indicator variable for any initial inpatient spending. "Initial" spending is defined as spending in the first three months of the plan for all covered members of the plan. $N = 102,022$.

Appendix Table A3: Additional Robustness Exercises

Specification	N	Log Initial Spending		Log Time to First Claim	
		Coeff on fp	(S.E.)	Coeff on fp	(S.E.)
(1) Baseline	102,022	-1.08	(0.29)	0.357	(0.114)
Panel A: Alternative sets of fixed effects					
(2) Don't limit to within firm	102,022	-1.07	(0.30)	0.320	(0.121)
(3) Don't control for tier	102,022	-3.98	(0.76)	1.943	(0.373)
(4) Tier x firm interactions	102,022	-1.04	(0.29)	0.355	(0.114)
Panel B: Family vs. single tier					
(5) Family tier	43,358	-0.90	(0.42)	0.132	(0.124)
(6) Single tier	58,664	-1.15	(0.40)	0.579	(0.193)
Panel C: Using additional plan characteristics to construct mp					
(7) Baseline (Firms A and C)	54,710	-0.81	(0.32)	0.263	(0.127)
(8) Firms A and C, refined fp measure	54,710	-0.90	(0.36)	0.293	(0.141)

Table reports results from alternative analyses of the relationship between initial medical utilization and expected end of year marginal price. The first row shows the baseline results (see last row in Table 5) from estimating equation 8 which pools the data across firms. In addition to the expected end of year marginal price, the regressions also include plan by coverage tier fixed effects, join month fixed effects and firm by month fixed effects. Standard errors are clustered on join month by coverage tier by firm. Alternative rows report single deviations from this baseline. In Row 2 we remove the firm by join month fixed effects from the baseline. In Row 3 we remove the controls for coverage tier (so that there are plan fixed effects but not plan by coverage tier fixed effects) from the baseline. In row 4 we add firm by coverage tier fixed effects and firm by coverage tier by join month fixed effects to the baseline. In rows 5 and 6 we stratify the sample by coverage tier. In Panel C we limit the analysis to the two firms (Alcoa and Firm C) in which we observe the in-network coinsurance rate for all plans (see Appendix Table A1 for details). Row 7 reports the baseline results limited to those two firms; Row 8 shows the sensitivity to using a refined measure of future price which accounts for the coinsurance rate (see Appendix Table A4 for details). As expected, not accounting for the coinsurance rate in our baseline future price measure (row 7) biases downward our estimated impact of the future price (compare rows 7 and 8).

Appendix Table A4: Alternative construction of future price

Employer	Plan	Expected end-of-year price[a]			Refined expected end-of-year price[b]		
		Joined plan in:			Joined plan in:		
		Feb-Apr	May-Jul	Aug-Oct	Feb-Apr	May-Jul	Aug-Oct
Alcoa	A0	0.000	0.000	0.000	0.0994	0.0995	0.0997
	A1	0.512	0.603	0.775	0.560	0.643	0.798
Firm C	C0	0.000	0.000	0.000	0.000	0.000	0.000
	C1-C3	0.543	0.633	0.811	0.589	0.670	0.830

[a] Expected end-of-year price is equal to the fraction of individuals who do not hit the deductible by the end of the calendar year (and therefore face a marginal price of 1). It is computed based on the plan's deductible level(s), join month, and the annual spending of all the employees in one's plan and join month; we compute it separately for family and single coverage within a plan and report the enrollment-weighted average.

[b] "Refined" expected end-of-year price is equal to the coinsurance rate times the fraction of individuals who hit the deductible but not the out-of-pocket maximum by the end of the year (and therefore face a marginal price equal to the coinsurance rate) + the fraction of individuals who do not hit the deductible by the end of the calendar year (and therefore face a marginal price of 1.) The refined expected end-of-year price is computed in the same manner as described above for the expected end-of-year price.

Appendix Table A5: Differences in observables by plan and join month

Employer	Plan	Deductible	Indicator for Old (>=45)		Indicator for Female	
		(Single/Family)	Difference	DD	Difference	DD
		[N = enrollees]	(1)	(2)	(3)	(4)
Alcoa	A0	0	-0.009		-0.011	
		[N = 3,269]	(0.004)		(0.003)	
	A1	250/500	-0.008	0.0020	-0.002	0.009
		[N = 3,542]	(0.002)	(0.0041)	(0.003)	(0.004)
Firm B	B0	0	-0.004		-0.003	
		[N = 37,759]	(0.003)		(0.002)	
	B1	150/300	-0.010	-0.0059	-0.004	-0.001
		[N = 9,553]	(0.004)	(0.0026)	(0.004)	(0.003)
Firm C	C0	0	-0.014		0.009	
		[N = 27,968]	(0.002)		(0.002)	
	C1-C3	200-300/500-750	-0.019	-0.0045	0.009	0.000
		[N = 19,931]	(0.003)	(0.0032)	(0.003)	(0.003)

Table reports coefficients (and standard errors in parentheses) from regressing the dependent variable on join month (which ranges from 2 (February) to 10 (October)). The dependent variables are demographic characteristics (defined in the top row) with overall means for "old" (i.e. age 45+) of 0.27 and for female of 0.48. Columns (1) and (3) report the coefficient on join month separately for each plan, based on estimating equation 4; the regressions also include an indicator variable for coverage tier (single vs. family). Columns (2) and (4) report the difference-in-differences coefficient on the interaction of join month and having a deductible plan, separately for each firm, based on estimating equation 5; the regressions also include plan by coverage tier fixed effects and join month fixed effects. Standard errors are clustered on join month by coverage tier.

Appendix Table A6: Summary statistics of the Rand data

Coinsurance rate	Maximum Dollar Expenditure (MDE)	Number of family years (Number of families in year 1)	Average MDE (Adjusted[a])	Share of family years who hit the MDE	Expected end-of year price[b]	Avg. time to first claim (Days)[c]	Effect of end-of-year price on Log(Time to First Claim)[d]	
							OLS	Tobit
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Plan-by-plan analysis								
100%	5% of income up to $1,000	33 (33)	$533	0.33	0.67	70	-2.94	-2.66
	10% of income up to $1,000	29 (29)	$801	0.21	0.79	82	(2.65)	(2.84)
	15% of income up to $1,000	33 (33)	$794	0.21	0.79	64		
95%	5% of income up to $1,000	418 (84)	$559	0.40	0.57	88	-0.10	-0.70
	10% of income up to $1,000	342 (80)	$746	0.34	0.63	86	(2.19)	(2.41)
	15% of income up to $1,000	470 (101)	$817	0.33	0.63	99		
50%	5% of income up to $1,000	111 (26)	$535	0.28	0.36	58	4.16	5.30
	10% of income up to $1,000	76 (17)	$779	0.16	0.42	84	(4.98)	(5.25)
	15% of income up to $1,000	308 (84)	$847	0.19	0.40	80		
50% for dental and mental health; 25% for all other	5% of income up to $750	189 (41)	$499	0.28	0.22	78	4.77	4.89
	10% of income up to $750	226 (44)	$584	0.31	0.22	59	(3.82)	(3.80)
	15% of income up to $750	159 (30)	$689	0.16	0.26	62		
	5% of income up to $1,000	18 (18)	$523	0.28	0.23	27		
	10% of income up to $1,000	19 (19)	$600	0.16	0.26	40		
	15% of income up to $1,000	13 (13)	$837	0.08	0.29	65		
25%	5% of income up to $750	192 (22)	$518	0.17	0.21	73	0.09	-2.26
	10% of income up to $750	208 (31)	$617	0.17	0.21	61	(21.71)	(22.15)
	15% of income up to $750	207 (26)	$683	0.18	0.21	61		
	5% of income up to $1,000	86 (52)	$535	0.14	0.22	71		
	10% of income up to $1,000	70 (43)	$818	0.11	0.22	38		
	15% of income up to $1,000	70 (44)	$816	0.16	0.21	37		
0%	–	2,376 (620)	–	1.00	0.00	46		
Panel B: Pooling across plans								
All positive coins plans, with coins dummies		3,277 (870)					-0.36	-0.08
							(1.51)	(1.64)
All positive coins plans, pooled		3,277 (870)					0.52	0.73
							(1.12)	(1.23)
All plans, pooled		5,653 (1,490)					1.90	1.96
							(0.83)	(0.90)

[a] Regression adjusted for differences in site, start month, and year across plans (see Newhouse et al. (1993, Appendix B) for more details).

[b] Expected end-of-year price equals the share of families not hitting the MDE (in the given plan) times the coinsurance rate. For the mixed coinsurance rates plans, we weight the two coinsurance rates based on their shares of initial claims in the full sample; 25% of initial claims are for mental/dental.

[c] For families with no claims in a given year, time to first claim is coded as 367.

[d] Columns (8) and (9) show the coefficient on the expected end-of-year price fp from estimating variants of equation 9. In Panel A we regress log time-to- first-claim on the expected end-of-year price (see column (6)) and site and enrollment month by year dummies; plan assignment in the RAND experiment was random conditional on the location (site) and when the family enrolled in the experiment (see Newhouse et al. (1993, Appendix B) for more details). In Panel B we pool across plans and therefore add additional controls in the form of either coinsurance dummies (first row) or the coinsurance level directly (bottom two rows); the final row adds in the free care (0% coinsurance) plan. Standard errors are clustered on family.

Appendix Table A7: Approximating the response to a linear contract in the RAND data

Regressor	Dependent Variable					
	Log Initial Spending[a]		Log Annual Spending[b]		Log Time to First Claim[c]	
	OLS	IV	OLS	IV	Tobit	Tobit IV
Coins rate	-1.21	-1.19	-1.78	-1.65	0.88	0.51
	(0.73)	(1.03)	(0.73)	(1.03)	(0.55)	0.49)
Share hitting MDE	0.45	0.43	0.20	0.21	-0.23	-0.25
	(0.21)	(0.25)	(0.20)	(0.24)	(0.15)	(0.12)
Coins rate*	0.58	0.52	1.76	1.40	-0.54	0.14
Share hitting MDE	(1.79)	(2.53)	(1.78)	(2.53)	(1.32)	(1.21)

Sample consists of 5,653 family-years (1,490 unique families) in the RAND data in one of the positive coinsurance plans or the free care plan. "Share hit MDE" is the share of families in a given coinsurance and maximum dollar expenditure (MDE) plan who spend past the MDE during the year. Because plan assignment in the RAND experiment was random only conditional on site and month of enrollment in the experiment, all regressions control for site and start month fixed effects (see Newhouse et al. (1993, Appendix B) for more details). All regressions cluster standard errors on the family, except for the Tobit IV specifications, which is estimated using a minimum distance estimator (Newey, 1987). In the IV specifications, we instrument for the share of families in a given coinsurance and MDE plan who hit the MDE with the "simulated" share hitting the MDE; the "simulated" share is calculated as the share of the full ($N = 5,653$) sample which, given their observed spending, would have hit the MDE if (counterfactually) assigned to the given plan; the coefficient on the instrument in the first stage is 1.05 (standard error 0.003); the F-statistic on the instrument is 120,000. Appendix Table A6 provides more details on the plans in the RAND experiment, the distribution of the sample across the different plans, and the share of families who hit the MDE in each plan.

[a] Dependent variable is $log(s + 1)$ where s is the total medical spending of the employee and any covered family members in their first three months in the plan.

[b] Dependent variable is $log(s + 1)$ where s is the total medical spending of the employee and any covered family members in their full year in the plan.

[c] Dependent variable is $log(time)$ where "*time*" is the number of days to first claim by any covered family member, censored at 367 days

Appendix Table A8: Fit of the calibration exercise of medical events

	Total Spending		Inpatient		Outpatient	
	Actual	Simulated	Actual	Simulated	Actual	Simulated
Mean of log spending	6.57	6.53	2.08	1.61	6.18	6.06
Standard deviation of log spending	2.17	2.10	3.58	3.28	1.96	2.04
Share with any spending	93.7%	93.8%	25.8%	19.7%	93.5%	92.5%

The table reports summary statistics of the actual and simulated moments of the spending distribution for the RAND "free care" plan. Log spending is computed as log(spending+1) to avoid missing values. Simulated data are generated as described in Appendix B.

Appendix Table A9: Calibration and fit of the "primitive" price elasticity parameters

Imputed Value of δ	Value of obj. function at optimum	Value of α at optimum	Value of b at optimmum	Implied $E(\omega)$
0	19.9	0.30	0.20	0.60
0.1	10.1	0.25	0.20	0.56
0.2	15.5	0.25	0.20	0.56
0.3	11.6	0.30	0.25	0.55
0.4	11.9	0.30	0.25	0.55
0.5	14.0	0.35	0.30	0.54
0.6	16.6	0.35	0.30	0.54
0.7	24.5	0.35	0.30	54
0.8	34.6	0.35	0.35	0.50
0.9	28.7	0.35	0.35	0.50
1	29.7	0.35	0.35	0.50

Imposed	Value of	Mean log spending			Std. log spending			Share with zero spending		
δ	obj. function	Coins 25%	Coins 50%	Coins 95%	Coins 25%	Coins 50%	Coins 95%	Coins 25%	Coins 50%	Coins 95%
Actual (observed) moments		6.08	6.04	5.53	2.36	2.35	2.71	90.7%	90.7%	85.2%
0	113.9	6.09	5.92	5.28	2.28	2.39	2.81	91.8%	90.6%	84.0%
0.1	45.3	6.09	5.93	5.39	2.28	2.39	2.77	91.8%	90.6%	84.8%
0.2	20.1	6.09	5.94	5.46	2.28	2.39	2.75	91.8%	90.6%	85.3%
0.3	11.6	6.10	5.95	5.52	2.28	2.39	2.73	91.8%	90.7%	85.7%
0.4	11.9	6.10	5.95	5.57	2.28	2.39	2.72	91.8%	90.7%	86.0%
0.5	17.7	6.10	5.96	5.61	2.28	2.39	2.71	91.8%	90.7%	86.3%
0.6	27.6	6.10	5.97	5.65	2.28	2.39	2.70	91.8%	90.7%	86.5%
0.7	40.3	6.11	5.98	5.68	2.28	2.38	2.69	91.8%	90.7%	86.7%
0.8	55.8	6.11	5.99	5.72	2.28	2.38	2.68	91.9%	90.8%	86.9%
0.9	74.1	6.11	6.00	5.75	2.28	2.38	2.67	91.9%	90.8%	87.1%
1	97.5	6.11	6.01	4.79	2.28	2.38	2.67	91.9%	90.8%	87.2%

The top panel reports the values of a and b that minimize the objective function for different values of δ. The bottom panel reports goodness of fit measures for our choice of $a = 0.3$ and $b = 0.25$ for different values of δ. Log spending is computed as log(spending+1) to avoid missing values. Simulated data are generated as described in Appendix B.

Chapter 3

The Impact of Financial Aid on College Enrollment and Completion: Evidence From a Discontinuity in Federal Financial Aid Rules[1]

3.1 Introduction

Economists have long been interested in the effect of financial aid on college enrollment. The topic has obvious policy relevance, given that the federal and state governments provide tens of billions of dollars of financial aid each year with the goal of expanding access to college. In addition, understanding the effect of aid on enrollment can help answer the question of whether credit market failures are significantly reducing college enrollment rates. In theory, students who expect high returns from continued education should be able to borrow to pay for it, and so students should never fail to go to college for lack of funds. In practice, private credit markets may be incomplete due to adverse selection, moral hazard, and other factors, and, while the federal student loan program exists in part to remedy this market failure, federal loans are available only up to specified limits. Thus, it is possible that students are liquidity constrained.

There is a large literature on the effects of grant aid on college enrollment, which has converged on the estimate that enrollment increases by about 3-4 percentage points per $1,000 increase in available grant aid (see for example Dynarski (2002), Dynarski (2003), and Kane (2003); for a similar estimate for older students, see Seftor and Turner (2002)). There is also some evidence that additional grant aid leads students to choose more expensive colleges than they otherwise would,

[1] I am grateful to Amy Finkelstein and David Autor for extensive feedback on this project. I also thank Joshua Angrist, Sarah Cohodes, Daniel Fetter, Matthew Fiedler, Heidi Williams, and participants in the MIT Labor Lunch for helpful comments and suggestions and Leila Agha, Dan Feenberg, Brad Larson, and Amanda Pallais for help accessing data. Financial support from the National Science Foundation Graduate Fellowship is also gratefully acknowledged.

in particular, to choose private over public colleges (Kane, 2003). Because grant aid both provides additional liquidity and lowers the price of college, examining the effect of additional grant aid on enrollment is not a perfect test for liquidity constraints. However, the large responses researchers have found to fairly modest reductions in the price of college are at least suggestive of credit market failures.

Somewhat surprisingly, there is much less evidence on the effect of financial aid on persistence and completion, topics of increasing policy importance.[2] Over the past several decades, college participation rates in the United States have risen rapidly, but growth in degree attainment has stalled. Since the late 1960s, the share of 23-year-olds with any college education has increased by almost 30 percent, but the share with a B.A. has remained roughly flat, implying a sharp decline in completion rates (Turner, 2004). Among the most recent cohorts, degree attainment rates are quite low. Only 58% of students who started college at four-year institutions in 2003-2004 obtained a B.A. within six years; among students starting at two-year community colleges, only 35% obtained any degree within six years (Hunt-White, 2010).

Policymakers have become increasingly focused on the large number of students who start college but never finish. President Obama drew attention to low college completion rates during a major education policy address in 2010, and, in March of 2011, Secretary of Education Arne Duncan announced new incentives for colleges to raise graduation rates (Obama, 2010; Lewin, 2011). The Gates Foundation, the Lumina Foundation, and the American Association of Community Colleges have also all introduced major college completion initiatives within the past few years.

College completion rates are especially low for students from low-income families, and, when asked, many non-completers cite financial pressure as a reason they dropped out of college (Hunt-White, 2010). Moreover, over the same period in which college completion rates have fallen, college costs have risen rapidly, the share of costs covered by federal financial aid programs has gone down, and the share of students attempting to work part- or full-time while in college has gone up (Turner, 2004). Thus, it seems plausible that financial stress might be part of the reason for low and falling college completion rates and that additional financial aid might increase persistence. In addition, from a policy perspective, financial aid is a relatively simple lever to manipulate, and so it would be useful to know whether and how much increases in financial aid could contribute to raising completion rates.

As with enrollment, evidence on the effect of financial aid on completion is also important because it sheds light on whether credit market failures are responsible for a meaningful share of the dropout rate. Some have suggested that high college dropout rates are actually the result of an optimal learning process in which many students "try" college to see whether it passes a cost benefit test, learn things about themselves or the college experience that show it does not, and then (appropriately) leave (see for example Strange (2009)). On the other hand, if college completion rates are sensitive to modest increases in financial aid, then it seems more likely that students are

[2] There is a larger literature on the effects of counseling and remediation services on persistence. See for example Bettinger and Baker (2011) and Calcagno and Long (2008).

dropping out for lack of liquidity.

Recently, there has been some very interesting experimental work on the effect of performance-contingent financial aid on persistence. In an experimental intervention focused on low-income single mothers at two Louisiana community colleges, Richburg-Hayes et al. (2009) find that offering $1,000 per-semester performance-based scholarships, along with various counseling services, increases persistence from year one to year two by 10-15 percentage points. A similar intervention focused on traditional-age college students at a large Canadian University found smaller but positive effects on persistence for women, though no effect for men (Angrist et al., 2009).

In this paper, I estimate the effect of non-performance-based aid on persistence for a nationally representative sample of non-traditional age college students. Understanding the effect of non-performance-based aid is valuable both because the overwhelming majority of financial aid is not based on performance and because the response to non-performance-based aid is closer to a pure liquidity effect. Put differently, by estimating the effect of non-performance-based aid on persistence and completion, I can isolate the effect of relaxing financial constraints, as opposed to the combined effect of relaxing financial constraints and creating new incentives.

My basic strategy is to exploit a discontinuity in federal financial aid rules that allows 24-year-old students to qualify for substantially more federal financial aid than 23-year-olds. Specifically, I take advantage of the fact that, for purposes of federal financial aid, undergraduate students who are not married and do not have children are classified as "dependent" or "independent" based on whether they have turned 24 by January 1 of the "award year" (the academic year for which they are receiving financial aid). Independent students qualify for additional grant aid and are eligible to take out much larger federal loans.

The discontinuity in financial aid at age 24 creates a quasi-experiment for looking at the effects of financial aid on persistence. Consider two students who both enter college at age 22; one turns 23 on December 31, while the other turns 23 on January 1. Because of this trivial difference, the first student may qualify for thousands of dollars more aid for her second year of college. By examining persistence and completion rates by date of birth, I should thus be able to estimate the effect of aid on persistence and completion.

In addition to examining the effect of aid on persistence, I also estimate the effect on total enrollment at different types of institutions. Most of the previous research on grant aid and enrollment predates the explosive growth of for-profit institutions. Using more recent data, I find suggestive evidence that enrollment at these institutions may be especially sensitive to financial aid. For-profit colleges have recently come under intense scrunity from policymakers who are concerned about the quality of education they provide as well as about the costs they impose on the federal financial aid system. Thus, understanding the effects of federal financial aid on enrollment at for-profit institutions is particularly relevant to current policy debates.

Because of the nature of the quasi-experiment, my results will be most relevant to older, "non-traditional age" college students. While these students are a minority of U.S. undergraduates, they are a sizable and growing group. Roughly five million college students, or about 25% of all

U.S. undergraduates, are between 23 and 30 years old; this percentage is much higher than it was a few decades ago and is projected to rise over the next ten years. Older students drop out of college at much higher rates than traditional-age students, and they have been much less studied (Hunt-White, 2010). They are also a major constituency of for-profit colleges.

The remainder of the paper proceeds as follows. Section 2 provides background on the federal financial aid system and additional detail about the quasi-experiment. Section 3 outlines my empirical framework, and section 4 describes my data. Section 5 analyzes the increase in aid at age 24, section 6 presents results on enrollment and school choice, and section 7 presents results on persistence and completion. Section 8 concludes.

3.2 Background on the Federal Financial Aid System

The federal government is a major source of financial assistance for college students, providing about \$33 billion in grant aid and originating more than \$100 billion in students loans each year. The first step in determining a student's eligibility for federal financial aid is to classify her as dependent or independent (of her parents). Students are classified as independent if they are married, have children or other dependents, are veterans or orphans or wards of the court, or - crucially for my purposes - if they are 24-years-old as of January 1 of the award year. Differences in how dependent and independent students are treated for purposes of awarding both grants and loans create my quasi-experiment.

3.2.1 Federal Grants

Federal grants are means-tested, and so they are awarded based on information provided on the Free Application for Federal Student Aid (FAFSA), a form used to elicit detailed information on income and assets. The information provided on the FAFSA is plugged into what is effectively a tax schedule, yielding the student's expected family contribution (EFC). The EFC represents what the federal formula implies the student and her parents can afford to pay for college.

The formula used to compute the EFC from total income and total assets is the same for dependent and independent students, but independent students do not have to report their parents' income or assets on the FAFSA. That is,

$$EFC_{dep} = f(income_{parents} + income_{student}, assets_{parents} + assets_{student}),$$

while

$$EFC_{ind} = f(income_{student}, asset_{student}).$$

So the EFC for a given student if she is classified as independent will be weakly less than her EFC if she is classified as dependent.[3]

[3] This is not quite strictly true, because independent students can qualify for an automatic \$0 EFC if their parents' incomes are low enough, irrespective of their own incomes. Hence, a high-income student with low-income parents

As a result, a student's federal grants will be weakly larger if she is classified as independent. Pell Grants, the most important federal grant program, are calculated as cost of attendance less EFC, up to some maximum (currently $5,550), so when a student's EFC goes down, her Pell Grant goes up as much as one for one.[4] Other, smaller federal grant programs have additional criteria besides financial need, but they all require that students be need-eligible for Pell Grants. Thus, a student's chance of qualifying for other federal grants is also (weakly) higher if she is classified as independent. Additionally, most state grant programs and some institutional grant programs piggyback on the federal EFC calculation, and so being classified as independent may also increase a student's state and institutional grant awards.

3.2.2 Federal Loans

The largest federal loan program, the Stafford Loan Program, is not means-tested, but independent students are still eligible for larger loans. Specifically, Stafford loans are available to all students, irrespective of financial need, credit history, or other criteria, but only up to specified annual limits (shown in Table 3.1) that depend on a student's year in school and on whether she is classified as dependent or independent. Independent students are eligible to take out $4,000-$5,000 more in Stafford Loans each year. (To put these and other figures in this paper in context, average tuition and fees at two-year public colleges in 2010-2011 were about $2,700, average tuition and fees at four-year public colleges were about $7,600, average tuition and fees at for-profit colleges were about $13,900, and average tuition and fees at non-profit private colleges were about $27,300 (College Board, 2010). So while the higher loan limits available to independent students would not put much of a dent in the difference in cost between attending a private non-profit and a public four-year college, they come close to covering the difference in tuition between community colleges and four-year public colleges or between four-year public colleges and for-profit colleges.)

Independent students are also more likely to qualify for other federal loans with lower interest rates. The interest rate on Stafford Loans is currently 6.8%, but students with high financial need (as determined by their EFC) can qualify for "subsidized Stafford Loans," with an interest rate of 4.5%, or Perkins Loans, with an interest rate of 5%. Since a given student will have a lower EFC if classified as independent, being classified as independent makes her more likely to qualify for these lower-interest rate loans. (The limits on annual borrowing apply to the sum of a student's subsidized and unsubsidized Stafford Loans, though they exclude Perkins Loans, which are much less prevalent.)

The federal government also provides loans to college students' parents through the PLUS Loan Program. This could be a source of nonmonotonicity in my quasi-experiment, since PLUS Loans are available to parents of dependent but not independent students, and they are limited only by

could actually be better off if classified as dependent. However, at most 2.5% of students fall into this category, and probably less.

[4] "Cost of attendance" is an estimate provided by the student's school that includes not only tuition but also other fees, books, and room and board. Hence, a student's Pell Grant can exceed her tuition.

the student's total cost of attendance. Thus, in theory, a student's total borrowing capacity could actually fall at age 24 due to the loss of access to the PLUS Loan Program. However, a student could only lose from being classified as independent if her parents would otherwise have borrowed more than $4.000 on her behalf. Since less than 3 percent of 23-year-old students' parents borrow this much through the PLUS Program, the nonmonotonicity problem is minimal.

3.3 Empirical Framework

My analysis is a "fuzzy regression discontinuity (RD)" design in which the running variable is age as of January 1 of the award year (with a cut-off at 24), the first stage is the effect of being classified as independent on financial aid, and the reduced form is the effect of being classified as independent on enrollment or persistence.

In this paper, I analyze the first stage and reduced form separately. Because being classified as independent increases both grant aid and available federal loans, it is difficult to summarize the first stage in a single number and therefore difficult to know what to use as the first stage in computing two-stage least squares (2SLS) estimates. Thus, it seems to make more sense to analyze the various components of the first stage and then analyze the reduced form.[5]

My basic specification is a local linear regression model with a bandwidth of 1.7 years, using a triangular kernel. That is, for any given first stage or reduced form outcome, I estimate the equation:

$$Outcome_i = \alpha + \beta_1 Over24_i + \beta_2 |Age_i - 24| + \beta_3 |Age_i - 24| * Over24_i + \varepsilon_i,$$

where i indexes students, the coefficient of interest is β_1, and observations are weighted according to $\frac{|age_i - 24|}{1.7}$.[6] The advantage of parameterizing the age variables as the absolute value of age less 24 is that, with this parametrization, α gives the "control group mean" (the estimate for students who are 23 and eleven months years-old), and so $\frac{\beta_1}{\alpha}$ gives the percent change in the outcome at age 24. Age is measured in months. I report robust standard errors.[7]

[5] Another concern is that, if there is an enrollment response, then the estimated first stage may in part reflect changes in the composition of the student population (due to the increase in enrollment). In principle, one might therefore prefer to compute the first stage by calculating how much more aid 23-year-old students would have received if they were 24, instead of by comparing 23-year-old and 24-year-old students. To try to evaluate how big a difference this would make, I take advantage of the fact that the EFC is close to a summary statistic for federal financial aid eligibility. I calculate EFCs for all students (below and above 24) as if they were classified as independent and look for a discontinuity in the average EFC at age 24. I find no evidence of such a discontinuity, suggesting that, even if there is an enrollment response to the increase in aid at age 24, it does not change the composition of the student population in such a way as to have much effect on the estimated first stage.

[6] For the enrollment outcomes, I regress total enrollment by age on the various age variables. So for these regressions, i indexes birth months, rather than students.

[7] Because I only have data on month of birth (not time of birth with unlimited precision), my running variable is effectively discrete. In such cases, Lee and Card (2008) recommend clustering standard errors on the running variable. None of my results would change if I reported clustered rather than robust standard errors, but the robust standard errors are generally more conservative (i.e. larger).

I chose a bandwidth of 1.7 years (i.e. including students age 22.3 to age 25.7) based on a combination of the Imbens-Kalyanaraman algorithm and "a priori" information (Imbens and Kalyanaraman, 2009). Applied to my data, the Imbens-Kalyanaraman algorithm yields optimal bandwidths ranging from one to two years depending on the particular outcome variable. I wanted to fix a single bandwidth across outcomes. I also thought it made sense to choose a bandwidth that would exclude most students who started college straight out of high school, since these "traditional" students would be expected to be quite different from the non-traditional age students affected by my instrument. If they started elementary school at the typical age and were never held back a grade, the oldest traditional students would be 22.25-years-old as of January 1 of their senior year of college, and so a bandwidth of 1.7 years just excludes them. However, I also experimented with bandwidths of 1, 1.5, and 2 years, and the results reported in this paper are robust to the choice of bandwidth.

I restrict my sample to students who are classified as dependent or independent based on age, and so I drop students who are married, have dependent children, or are veterans, orphans, or wards of the court. (Dropping the unaffected students should not affect the 2SLS estimates, but it does make the first stage and reduced form estimates more easily interpretable and may improve power.) In some cases, I also consider subgroups of students, such as students whose parents did not finish college or students attending a particular type of institution.

3.3.1 Possible Confounds

There are two possible confounds that could invalidate the interpretation of my results as reflecting the causal effect of financial aid. First, there might be other direct effects of age as of January 1. In particular, some states and school districts use age on January 1 as a cut-off for kindergarten entry, allowing students to enter kindergarten if they will be five-years-old by January 1 of their kindergarten year. Thus, students who are just-24 versus not-quite-24-years-old as of January 1 of a given academic year may have entered school with different cohorts and may be different numbers of years out of high school. These "cohort effects" might directly influence students' college-going or persistence decisions.

I attempt to address these possible cohort effects by including month of birth dummies in all regressions. If the cohort effects are roughly equal at the age 23, age 24, and age 25 cut-offs, then this should solve the problem. In practice, I find that including month of birth dummies in the regressions makes almost no difference to the results, which suggests that cohort effects may not be that important.

In addition, it turns out that, during the period in which students in my sample entered school, more states used a September 1 than a January 1 cut-off for kindergarten entry. Thus, if cohort effects were driving either the first stage or the reduced form results, one would expect to see even larger discontinuities between students who are just-24 versus not-quite-24 as of September 1. But there do not appear to be any discontinuities at this cut-off.

A second possible confounding influence would be anticipation effects. Suppose students are fully informed about the federal financial aid system and fully understand the discontinuity at age 24. In that case, 23-year-olds with January or later birthdays who are thinking about enrolling in college might decide to wait a year, potentially invalidating the comparison between just-24-year-old and not-quite-24-year-old college students.

In principle, one could test for anticipation effects by looking for a discontinuity in enrollment at age 23. Consider two students who are both 22-years-old in September of a given year and are contemplating enrolling in college; one will turn 23 in December and the other in January. The first student need delay enrollment only one year in order to be classified as an independent student, but the second student would have to delay for two years. Thus, assuming the cost of delay rises with the length of the delay, anticipation should produce a discontinuity in college enrollment at age 23.

While the estimates are not terribly precise, I do not find any evidence of a decrease in total enrollment or enrollment at any particular category of institution (including for-profits) at age 23.

3.4 Data

I implement my analysis using the restricted use versions of two datasets available through the National Center for Education Statistics of the Department of Education: the National Postsecondary Student Aid Study (NPSAS) and the Beginning Postsecondary Students Longitudinal Study (BPS).

The NPSAS, which is conducted every four years, surveys a representative sample of the U.S. undergraduate population. I pool the three most recent (and largest) waves of the NPSAS, the 1999-2000, 2003-2004, and 2007-2008 surveys.[8] Each of these surveys covers 50,000-100,000 students (roughly a 0.25%-0.5% sample of the full undergraduate population), yielding a total sample of about 243,000 students. From this, I construct a sample of 32,705 students who are between 22.3- and 25.7-years-old as of January 1 and who are potentially affected by the instrument (i.e. who are not married, do not have children, and are not veterans, orphans, or wards of the court). Fortunately, as shown in Table 3.1, the relevant federal student aid rules were quite stable over the 1999-2008 period, allowing me to pool survey waves without having to worry about major policy differences.[9]

The major virtue of the NPSAS is that it contains excellent administrative data on financial aid. Students in the NPSAS are matched to their FAFSAs and to their records in the federal grant and loan data systems, and so I have accurate and precise data on federal grant and loan receipt, as well as detailed data on income and assets.[10] I also have school-reported data on state and institutional grants and student-reported data on private loans. In addition, I know what school or schools each student attended, and I have a variety of demographic data for students, including month of birth as well as race, gender, and parents' educational attainment.

[8] Earlier waves of the NPSAS do not have administrative data on financial aid.

[9] I do adjust all financial aid variables for inflation, using the CPI-U. Dollar amounts are reported in 2010 dollars.

[10] Because the financial data come from the FAFSA, I have information on both students' and parents' income and assets for dependent students; for independent students, I have only the student's financial data.

The main deficiency of the NPSAS for my purposes is that it has no panel component. All I know is that students were enrolled in college during the academic year covered by a given NPSAS wave; I do not know whether they were enrolled during the previous or subsequent year. Hence, while I use the NPSAS to estimate the first stage and to estimate the effect of aid on total enrollment, I cannot use these data to estimate effects on persistence or completion.

For some waves of the NPSAS, however, there is a corresponding wave of the BPS that follows a sub-sample of the NPSAS first-time, first-year students for the following four or five years. I pool all available waves of the BPS, which follow the 1989-1990, 1995-1996, and 2003-2004 NPSAS cohorts. (Again, as shown in Table 3.1, the federal student aid rules were fairly stable over the relevant period.) The BPS provides detailed information on students' enrollment patterns and degree attainment during the follow-up period, and so it is highly appropriate for examining persistence and completion. It also provides basic demographic information, including month of birth.

The downside of the BPS is that is is extremely small. Each wave covers only 7,000-19,000 students total, and there are also substantial attrition problems. Thus, even pooling waves, I end up with a sample of only 1,231 first-year students who will be between 22.3- and 25.7-years-old as of January 1 of the next academic year, who are potentially affected by the instrument, and for whom I have second-year enrollment data. I also examine year-to-year persistence for a sample of all students who are between 22.3- and 25.7-years-old at any point during the BPS survey and who are still enrolled in school and have not yet completed a degree at that point. Even this approach, however, gives me only 2,238 observations for which I have next-year enrollment data.

3.4.1 Sample Summary Statistics

Table 3.2 provides summary statistics for my NPSAS sample and compares these students with the traditional-age (18 - 21-year-old) college students in the NPSAS. Table 3.3 provides summary statistics for my BPS first-year students sample and compares these students with traditional-age (18 - 19-year-old) first-year students in the BPS.

Based on observable characteristics, the students in the NPSAS sample actually do not look that difference from traditional-age college students. More of them are men (reflecting the sample restriction that drops married students and students with dependent children), but their family backgrounds seem similar to those of traditional-age students. They are more likely to be attending four-year colleges and are much less likely to be enrolled full-time.

In contrast, the students in the BPS sample do look quite different from traditional-age first-year students. More of them are black or Hispanic, and fewer of their parents went to college. They are much more likely to be attending community colleges or for-profit colleges (fully 30.5% attend for-profits) and much less likely to be attending four-year public or non-profit colleges. In addition, their persistence rates are far below those of traditional-age students.

It is not surprising that the students in the BPS first-year students sample come from less privileged backgrounds than the students in the NPSAS sample, since students who are just starting college at age 23 or 24 are more negatively selected than students who are simply still enrolled at age 23 or 24 (and the NPSAS sample includes both types of students, whereas the BPS sample includes only the former). It is potentially somewhat worrisome, however, that the sample I use to compute the first stage is so different from the sample I use to examine persistence. For this reason, I do not report "split sample" 2SLS results. If I had more power in the BPS, I would try to reweight my NPSAS first stage to match the characteristics of the BPS sample. However, as discussed below, my BPS results are so imprecise that this did not seem worth doing.

3.5 The Increase in Financial Aid at Age 24

3.5.1 Grants

Figures 3.1 and 3.2 and Table 3.4 show the increase in grant aid that occurs at age 24. In the graphs, each circle is an average for students born in a given month; the red, solid circles correspond to the 22.3 - 25.7-year-old students included in the regressions. As noted above, the regressions are parametrized such that the coefficient on the constant is the "control group mean," the average level of aid for students who are 23 and eleven months years-old.

As shown in the graphs and table, there is a sharp, roughly 20 percentage point increase at age 24 in the share of students receiving Pell Grants, as well as a sharp, roughly \$650 (90%) increase in the average Pell Grant per student. Since only 50% of 24-year-olds (and 30% of 23-year-olds) receive Pell Grants, this is perhaps better thought of as a roughly \$1,300 average increase in aid for about half of students, with no change for the remainder.

The increase in Pell Grants is amplified by a \$185 increase in other federal grants and a \$190 increase in state grants, presumably due to state programs that base their grant awards on the federal EFC.[11] There is no statistically significant change in institutional grant aid, implying that, on average, institutions neither raise their awards in response to the drop in the EFC at age 24 nor lower them to offset the increases in federal and state aid.[12]

Total grant aid per student increases by about \$1,100, or 55%, at age 24. Again, however, since only 60% of 24-year-olds (and a smaller fraction of 23-year-olds) receive any grant aid, this is perhaps best thought of as a roughly \$1,800 increase in average grant aid for 60% of students,

[11] In the other federal grants, state grants, and institutional grants graphs, there is a striking increase in aid from age 20 to 22, followed by a dramatic fall, with an apparent discontinuity at 22.25-years-old. In the case of other federal grants and state grants, the rise appears to be due to the fact that certain grants are only available to juniors and seniors. In the case of institutional grants, the increase is due to a compositional effect: a larger share of 20 - 22-year-olds are going to private non-profit institutions, which provide substantial institutional aid, while a smaller share are going to two-year public colleges, which provide very little.

The discontinuity at age 22.25 reflects the fact that the oldest "traditional" students (i.e. students who went straight from high school to college) are 22.25 on January 1 of their senior year of college. Thus, traditional students make up a meaningful share of all students to the left of age 22.25, but a negligible share to the right. Traditional students are more likely to be eligible for non-Pell federal grants and state grants and are more likely to attend institutions that award significant grant aid than are non-traditional students. Note that the regressions exclude students younger than 22.3-years-old.

[12] I had thought that both responses might be occurring and roughly offsetting each other, with institutional aid increasing at public institutions and falling at non-profits. It turns out, however, that there is no statistically significant change in institutional grants at any category of institution.

with no change for the remainder.

3.5.2 Loans

Figures 3.3 and 3.4 and Table 3.5 show the changes in borrowing that occur at age 24. Figure 3.3, and the first column of the table, show that 12% of students take advantage of the higher federal loan limits available to independent students. While about 15% of not-quite-24-year-olds have total federal loans above the dependent student loan limits (either because their parents take out PLUS loans on their behalf or because they are eligible for Perkins Loans), about 27% of just-24-year-olds do. This implies that about 12% of not-quite-24-year-olds are constrained by the federal loan limits (in spite of their parents' ability to borrow on their behalf), or, equivalently, that 12% of just-24-year-olds take out larger federal loans than they would have if they were not eligible for the independent student loan limits.

Figure 3.3 and the second column of Table 3.5 show that the higher federal loan limits appear to partially displace borrowing from private lenders but mostly increase total borrowing.[13] The increase in the share of students with total loans above the dependent student loan limits is 8 percentage points, less than the increase in the share of students with federal loans over the limits, but still quite substantial.

Figure 3.4 and the remaining columns of Table 3.5 show the changes in average federal, private, and total borrowing at age 24. These average changes are fairly small (and not statistically significant in the case of total borrowing), not surprising given that they are the result of increased borrowing by just the 12% of students who take advantage of the higher independent student loan limits. Moreover, some students may actually decrease their borrowing as a result of being classified as independent, if increases in their grant aid lead them to decrease their loans.

Thus, a more meaningful way to think about the changes in borrowing at age 24 may be to focus on the 12% of students who take advantage of the higher loan limits. The estimated increases in average federal and total borrowing imply that these students must be increasing their federal and total borrowing by an average of about \$3,750 and \$2,250 respectively. These figures are broadly consistent with the results of quantile regressions, which, while highly imprecise, suggest that borrowing at the 80^{th} to 90^{th} percentiles of the distribution of total borrowing increases by about \$2,000 at age 24.

As explained above, parents of dependent students are allowed to take out federal PLUS loans on their behalf. Thus, the fact that total federal borrowing increases sharply at age 24 indicates that there are younger students who would like to borrow more than the dependent student loan limits allow and whose parents will not borrow on their behalf. If the interest rates on PLUS loans and Stafford loans were the same, this result would imply a failure of contracting within the family, since, with perfect contracting, parents should be willing to take out PLUS loans which the students would then pay off. However, since the interest rate on PLUS loans is modestly higher

[13] By "private lenders," I mean lenders who are offering loans outside the federal student loan system, not private lenders who are originating and servicing federal student loans through the Federal Family Education Loan Program.

than the interest rate on Stafford loans. it could also be the case that the dependent students who would like to borrow more at the Stafford loan interest rates would not want to borrow more at the PLUS loan rates.

3.6 Effects on Total Enrollment and Enrollment By Type of Institution

Figure 3.5 plots the log of total NPSAS enrollment by age by type of institution, while Table 3.6 shows the estimated change in log enrollment at age 24.

These results are very imprecise. While the graph of log total enrollment, for example, looks fairly smooth around age 24, the 95% confidence interval from the regression includes both a decline in enrollment and an increase of more than 7%. The results for enrollment at two-year and four-year public colleges and private non-profit colleges are similarly uninformative.

The results for for-profit colleges are more interesting. The point estimate is marginally statistically significant at conventional levels ($p = 0.057$), and, visually, it looks like there may be a discontinuity in the graph. Moreover, the point estimate, 0.11 log points, is quite large.

Figure 3.6 and Table 3.7 show enrollment results limited to students whose parents do not have college degrees. Such students might be expected to be more sensitive to financial aid than the children of college graduates; they are also the main constituency of for-profit colleges.[14] While most of the results for the subgroup are just as uninformative as the results for the full sample (the point estimates are larger, but naturally even less precise), there is a clear discontinuity in the graph of for-profit enrollment at age 24. The table shows that the estimated increase in enrollment at for-profit institutions is 0.24 log points and is statistically significant at the 99% level.

These results provide evidence that enrollment at for-profit colleges, especially enrollment of students from lower-SES backgrounds, is quite sensitive to financial aid. Because of the general imprecision of the results, I cannot definitely compare the response at for-profits to the response at other types of institutions. Based on the point estimates, however, the response at for-profits appears to be considerably larger.

3.7 Effects on Persistence and Completion

Figure 3.7 and Table 3.8 show results for persistence and completion, using the BPS data. The first panel of Figure 3.7 plots the share of first-year students continuing to year two by their age as of January 1 of year two (the age that determines their eligibility for financial aid for the second year). The next panel plots the share of all students who have not yet completed a degree continuing on to the following year, again by their age as of January 1 of the following year. The final panel plots

[14] I restrict the sample based on parents' education rather than parents' income because I do not have data on parents' income for independent students.

the share of students obtaining a degree by the end of year five by their age as of January 1 of year two.[15]

Unfortunately, as is evident from the graphs and from the standard errors in the regression table, I do not have nearly enough power to draw any conclusions about persistence or completion. For example, the estimated effect on year one to year two persistence is small and negative (-0.6 percentage points). But the 95% confidence interval includes a 12 percentage point increase in retention, or about 11 percentage points per thousand dollars of grant aid.

3.8 Conclusion

This paper documents that the federal student aid rules lead to a large increase in aid at age 24. Grant aid increases by about $1,000 per-student, or 55%, at age 24, or by about $1,800 per affected student. About 12% of students take advantage of the higher federal loan limits available to independent students, and many of these students increase their federal borrowing by thousands of dollars.

Unfortunately, my data do not provide enough statistical power for me to reach conclusions about the effect of additional aid on persistence and completion. I do find evidence that enrollment at for-profit colleges is quite sensitive to federal financial aid, with an especially large enrollment increase among students whose parents are not college graduates. The fact that enrollment at for-profit institutions appears to be so sensitive to financial aid is especially interesting given that policymakers are currently quite concerned about whether these institutions are serving their students well and about the costs they impose on the federal student aid programs.

References

Angrist, Joshua, Daniel Land, and Philip Oreopoulos, "Incentives and Services for College Achievement: Evidence From a Randomized Trial," *American Economic Journal: Applied Economics*, 2009, *1*, 136–163.

Bettinger, Eric and Rachel Baker, "The Effects of Student Coaching in College: An Evaluation of a Randomized Experiment in Student Mentoring," *National Bureau of Economic Research Working Paper*, 2011.

Calcagno, Juan Carlos and Bridget Terry Long, "The Impact of Postsecondary Remediation Using a Regression Discontinuity Approach: Addressing Endogenous Sorting and Noncompliance," *National Center for Community College Research Working Paper*, 2008.

College Board, "Trends in College Pricing 2010," *Trends in Higher Education Series*, 2010.

[15] Unfortunately, for looking at effects on degree completion, I have to drop the 2003-2004 wave of the BPS, for which I currently have access to only the first two years of follow-up data. The remaining sample contains only 509 students.

Dynarski, Susan, "The Behavioral and Distributional Implications of Aid for College," *American Economic Review*, 2002, *92*, 279–285.

_ , "Does Aid Matter? Measuring the Effect of Student Aid on College Attendance and Completion," *American Economic Review*, 2003, *93*, 279–288.

Hunt-White, Tracy, *Persistence and Attainment of 2003-04 Beginning Postsecondary Students: After Six Years*, National Center for Education Statistics, 2010.

Imbens, Guido and Karthik Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *National Bureau of Economic Research Working Paper*, 2009.

Kane, Thomas J., "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going," *National Bureau of Economic Research Working Paper*, 2003.

Lee, David S. and David Card, "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 2008, *142*, 655–674.

Lewin, Tamar, "Incentives Offered to Raise College Graduation Rates," *New York Times*, 2011.

Obama, President Barack, "Remarks on Higher Education and the Economy at the University of Texas at Austin on August 9, 2010," 2010.

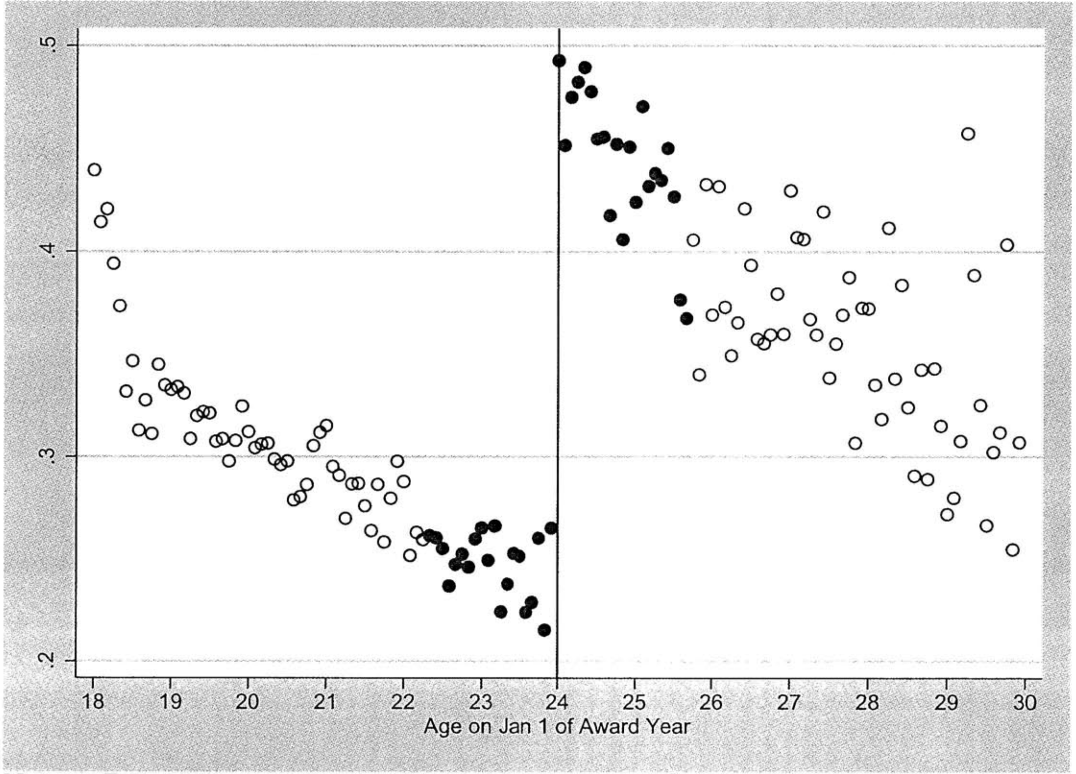
Richburg-Hayes, Lashawn, Thomas Brock, Allen LeBlanc, Christina Paxson, Cecilia Elena Rouse, and Lisa Barrow, "Rewarding Persistence: Effects of a Performance-Based Scholarship Program for Low-Income Parents," *MDRC Opening Doors Project*, 2009.

Seftor, Neil S. and Sarah E. Turner, "Back to School: Federal Student Aid Policy and Adult College Enrollment," *The Journal of Human Resources*, 2002, *37*, 336–352.

Strange, Kevin, "An Empirical Investigation of the Option Value of College Enrollment," 2009.

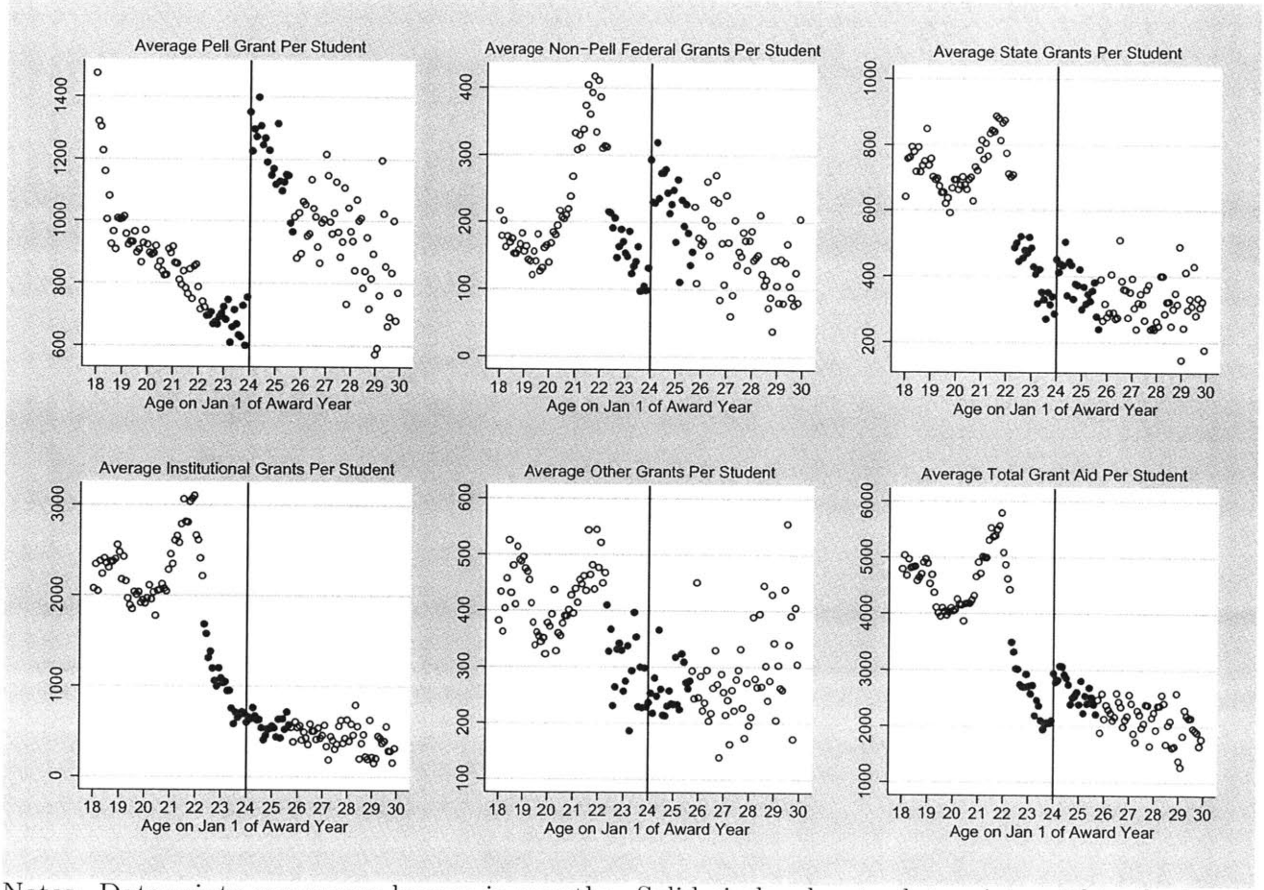
Turner, Sarah E., "Going to College and Finishing College: Explaining Different Educational Outcomes," in "College Choices: The Economics of Where to Go, When to Go, and How to Pay for It," National Bureau of Economic Research, 2004.

Figure 3.1: Share of Students Receiving Pell Grants by Age in Months



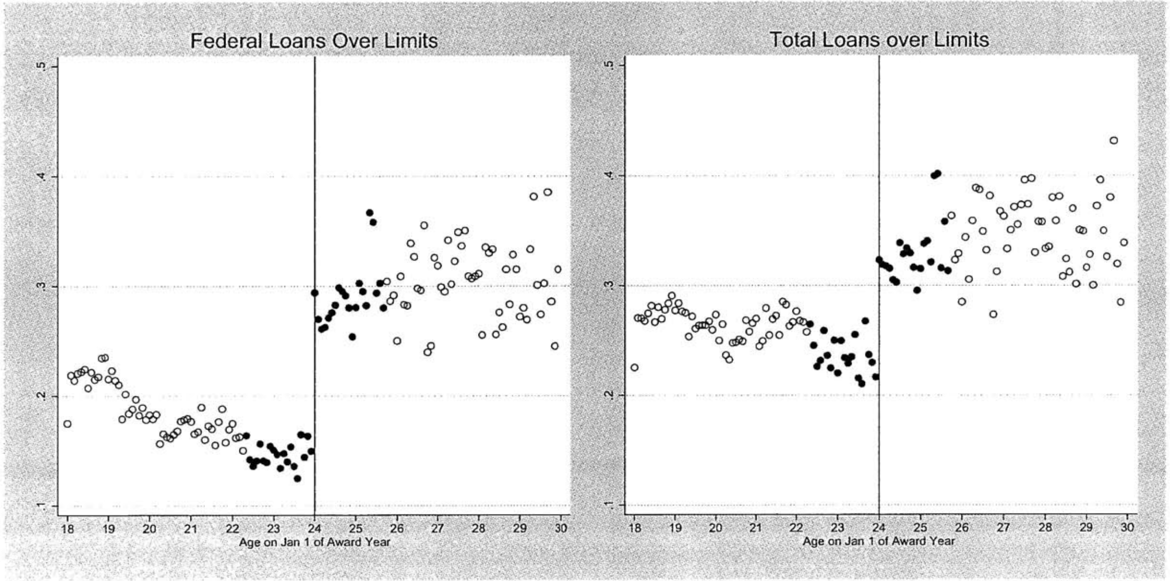
Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Figure 3.2: Average Grant Aid Per Student by Age in Months



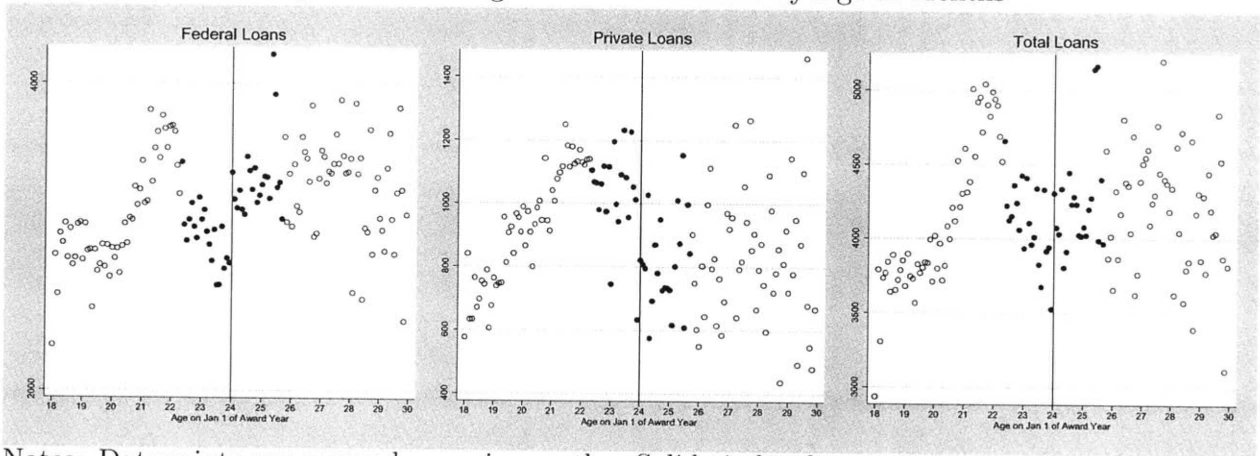
Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Figure 3.3: Share of Students with Loans Over Dependent Student Federal Limits by Age in Months



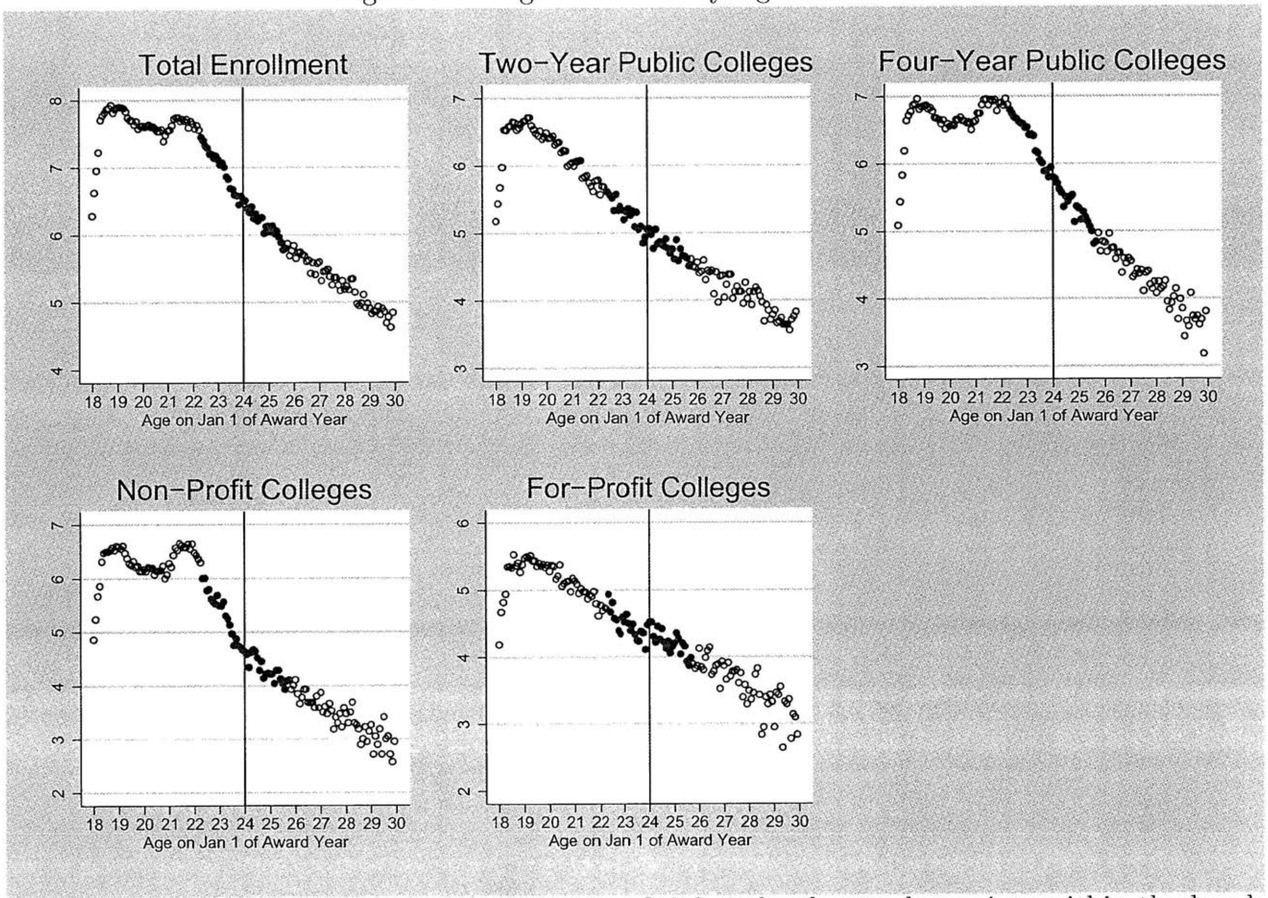
Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Figure 3.4: Average Loans Per Student by Age in Months



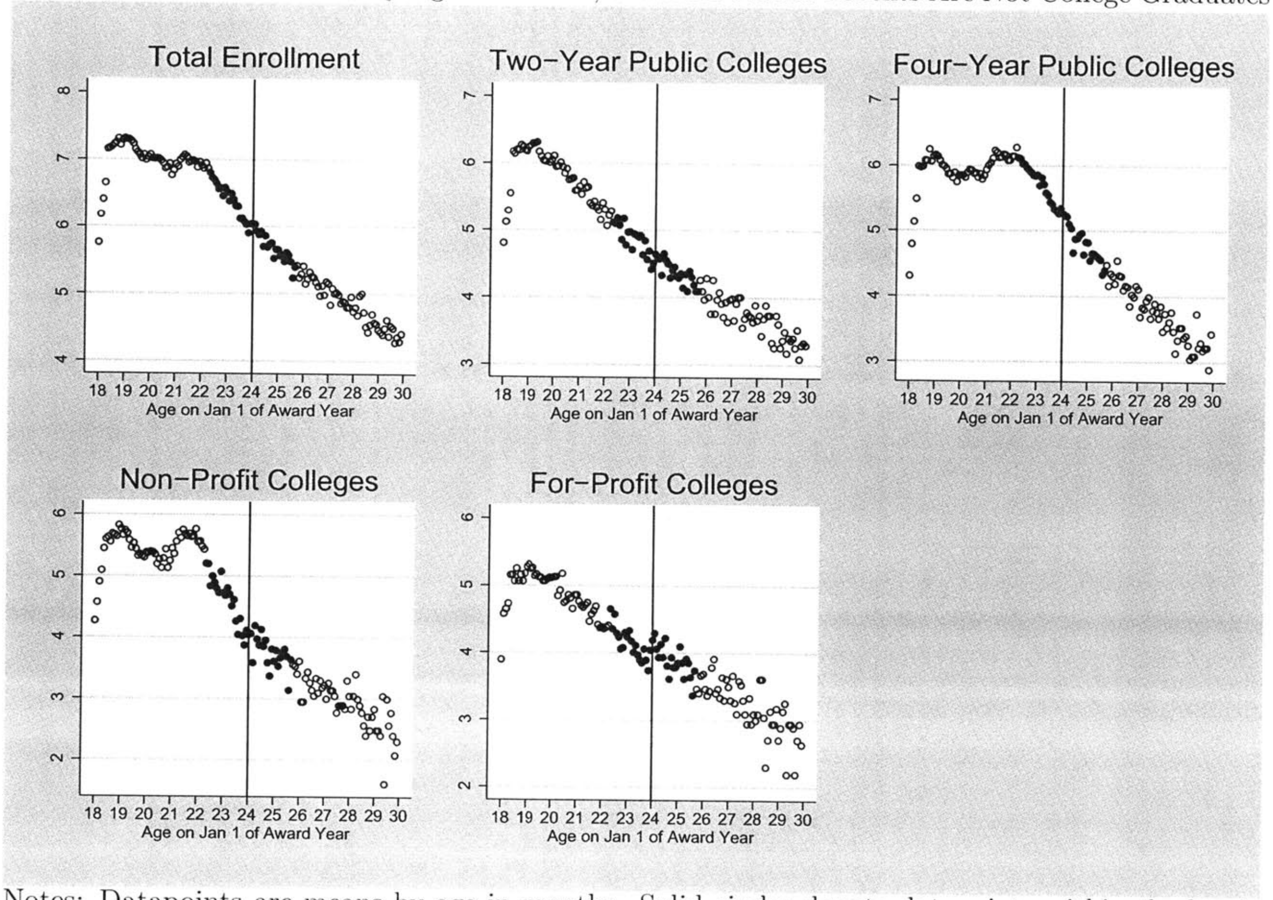
Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Figure 3.5: Log Enrollment by Age in Months



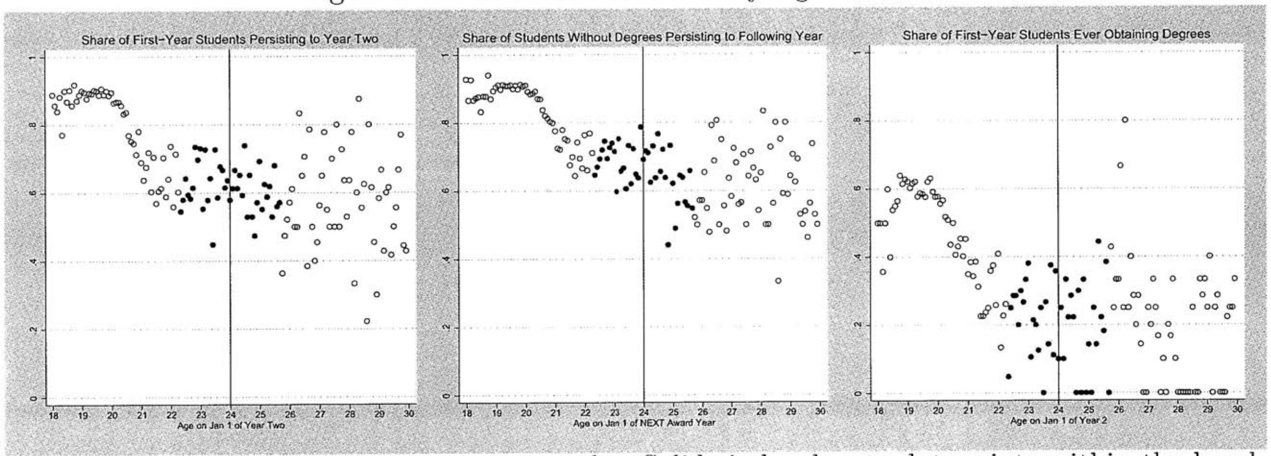
Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Figure 3.6: Log Enrollment by Age in Months, Students Whose Parents Are Not College Graduates



Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Figure 3.7: Persistence Measures by Age in Months



Notes: Datapoints are means by age in months. Solid circles denote datapoints within the bandwidth used for the regressions (1.7 years).

Table 3.1: Federal Financial Aid Parameters for Dataset Years

		Dependent Loan Limits			Independent Loan Limits		
	Max Pell Grant	Year 1	Year 2	Beyond	Year 1	Year 2	Beyond
Current Dollars							
1989-1990	2,300	2,625	2,625	4,000	6,625	6,625	8,000
1995-1996	2,340	2,625	2,625	5,500	6,625	7,500	10,500
1999-2000	3,125	2,625	2,625	5,500	6,625	7,500	10,500
2003-2004	4,050	2,625	2,625	5,500	6,625	7,500	10,500
2007-2008	4,310	3,500	4,500	5,500	7,500	8,500	10,500
2010 Dollars							
1989-1990	3,942	4,499	4,499	6,856	11,355	11,355	13,712
1995-1996	3,295	3,697	4,929	7,745	9,331	10,562	14,787
1999-2000	4,016	3,374	4,498	7,069	8,516	9,640	13,495
2003-2004	4,735	3,069	4,092	6,430	7,747	8,769	12,276
2007-2008	4,430	3,598	4,625	5,652	7,709	8,737	10,793

Table 3.2: Summary Statistics for NPSAS Sample

	Sample		Traditional Students	
	Percent	Std. Err.	Percent	Std. Err.
Female	49.9	0.3	57.7	0.2
White	74.6	0.2	73.5	0.1
Black	12.3	0.2	14.2	0.1
Hispanic	12.8	0.2	13.9	0.1
Other Race	7.1	0.1	6.5	0.1
Parent Any College	65.5	0.3	66.4	0.1
Parent BA	45.3	0.3	45.9	0.2
2-Year Public	20.5	0.2	26.8	0.1
4-Year Public	50.1	0.3	37.0	0.1
Non-Profit	18.9	0.2	25.8	0.1
For-Profit	9.9	0.2	10.4	0.1
Full-Time Student	54.0	0.3	70.7	0.1
N	32,705		109,689	

"Traditional students" are defined as those age 18-21 as of January 1 of the award year.

Table 3.3: Summary Statistics for BPS First-Year Students Sample

	Sample		Traditional Students	
	Percent	Std. Err.	Percent	Std. Err.
Female	43.0	14.	56.2	0.3
White	61.3	1.4	73.0	0.3
Black	16.9	1.1	10.3	0.2
Hispanic	14.4	1.0	10.7	0.2
Other Race	7.8	0.8	6.3	0.2
Parent Any College	44.3	1.5	67.4	0.3
Parent BA	24.0	1.3	48.3	0.3
2-Year Public	41.8	1.4	19.9	0.3
4-Year Public	13.8	1.0	38.1	0.3
Non-Profit	13.9	1.0	34.6	0.3
For-Profit	30.5	1.3	7.3	0.2
Full-Time Student	45.4	1.5	76.0	0.3
Persist to Year Two	62.1	1.4	87.6	0.2
N	1,231		25,383	

"Traditional students" are defined as those age 18-21 as of January 1 of the award year.

Table 3.4: Effects on Grant Aid

	Pell	Other Federal	State	Institutional	Other	Total
Over24	653.3***	186.2***	191.4***	104.8	-7.430	1128.3***
	(40.38)	(19.99)	(30.16)	(67.21)	(34.16)	(109.4)
Age	12.89	65.97***	171.7***	519.5***	36.74	806.8***
	(25.19)	(12.07)	(20.68)	(50.99)	(24.66)	(76.65)
Age*Over24	-156.6***	-129.3***	-264.3***	-633.9***	-37.14	-1221.3***
	(46.28)	(22.36)	(34.31)	(75.98)	(38.06)	(123.9)
Cons	717.4***	118.1***	302.5***	637.2***	252.2***	2027.4***
	(40.44)	(19.61)	(28.63)	(75.68)	(31.64)	(111.1)
N	32705	32705	32705	32705	32705	32705

Bandwidth is 1.7 years.

Robust standard errors in parentheses.

All regressions include month of birth dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.5: Effects on Borrowing

	Federal Loans Over Dep Limits	Total Loans Over Dep Limits	Federal Loans	Private Loans	Total Loans
Over24	0.120***	0.0822***	451.2***	-170.8*	274.2
	(0.0107)	(0.0117)	(115.2)	(80.68)	(152.0)
Age	-0.00502	0.00611	271.2**	58.57	342.5**
	(0.00648)	(0.00766)	(82.48)	(57.63)	(108.3)
Age*Over24	0.0301*	0.0130	-73.69	-82.79	-181.9
	(0.0125)	(0.0135)	(130.6)	(90.29)	(171.8)
Const	0.146***	0.222***	2830.2***	817.8***	3698.0***
	(0.0102)	(0.0117)	(124.7)	(75.40)	(154.6)
N	32705	32705	32705	32705	32705

Bandwidth is 1.7 years.

Robust standard errors in parentheses.

All regressions include month of birth dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.6: Effects on Log Enrollment by Type of Institution

	Total	Public 2-Year	Public 4-Year	Non-Profit	For-Profit
Over24	0.0161	0.0382	-0.0309	0.0292	0.110
	(0.0333)	(0.0576)	(0.0487)	(0.0612)	(0.0553)
Age	0.628***	0.398***	0.690***	0.894***	0.199***
	(0.0234)	(0.0388)	(0.0421)	(0.0514)	(0.0496)
Age*Over24	-0.990***	-0.654***	-1.153***	-1.267***	-0.418***
	(0.0359)	(0.0616)	(0.0545)	(0.0658)	(0.0645)
Cons	6.494***	4.911***	5.851***	4.605***	4.391***
	(0.0301)	(0.0225)	(0.0407)	(0.0360)	(0.0773)
N	41	41	41	41	41

Bandwidth is 1.7 years.

Robust standard errors in parentheses.

All regressions include month of birth dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.7: Effects on Log Total Enrollment by Type of Institution. Students Whose Parents Are Not College Graduates

	Total	Public 2-Year	Public 4-Year	Non-Profit	For-Profit
Over24	0.0451	0.0562	-0.0366	0.0597	0.237**
	(0.0366)	(0.0591)	(0.0553)	(0.0907)	(0.0739)
Age	0.535***	0.391***	0.566***	0.802***	0.297***
	(0.0199)	(0.0438)	(0.0321)	(0.0718)	(0.0594)
Age*Over24	-0.902***	-0.676***	-1.032***	-1.107***	-0.578***
	(0.0371)	(0.0574)	(0.0625)	(0.100)	(0.0712)
Cons	5.984***	4.488***	5.282***	4.085***	3.957***
	(0.0165)	(0.0273)	(0.0241)	(0.0477)	(0.0759)
N	41	41	41	41	41

Bandwidth is 1.7 years.

Robust standard errors in parentheses.

All regressions include month of birth dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.8: Effects on Persistence and Completion

	Persist to Year 2	Persist. Any Year	Complete Degree
Over24	-0.00632	0.0625	0.0211
	(0.0641)	(0.0460)	(0.0826)
Age	0.00465	0.0277	0.0335
	(0.0484)	(0.0356)	(0.0651)
Age*Over24	-0.0319	-0.137*	-0.0706
	(0.0717)	(0.0531)	(0.0917)
Cons	0.646***	0.704***	0.299**
	(0.0740)	(0.0515)	(0.102)
N	1231	2238	509

Bandwidth is 1.7 years.

Robust standard errors in parentheses.

All regressions include month of birth dummies.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$