2015

Essays on development economics and industrial organization

Young, Nathaniel

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

ESSAYS ON DEVELOPMENT ECONOMICS AND INDUSTRIAL ORGANIZATION

by

NATHANIEL YOUNG

B.A., Western Washington University, 2006
M.A., University of California at Davis, 2007

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2015

Approved by

First Reader

Dilip Mookherjee, PhD
Professor of Economics

Second Reader

Marc Rysman, PhD
Professor of Economics

Third Reader

Francesco Decarolis, PhD
Assistant Professor of Economics

Acknowledgments

I wish to thank the many people who have directly and indirectly supported me in the composition of this dissertation. I am profoundly grateful to my advisor, Dilip Mookherjee. He has greatly shaped my thinking and growth as an economist. I appreciate his generosity with his time and energy in discussing my ideas and offering advice. His passion for economics and boundless energy has provided an excellent example toward which to strive.

I also give special thanks to Marc Rysman, Francesco Decarolis and Samuel Bazzi. I thank Marc for his clear insights that have helped advance this work, and his encouragement in pursuing topics in IO and development. I am grateful to Francesco for our many long conversations, his thoughtful guidance and support. I thank Sam for his excellent comments and advice. Each has contributed greatly to my thinking in economics. I am very fortunate to have experienced the positive influence of these four individuals. I also wish to thank Kehinde Ajayi, Jordi Jaumandreu, Sergei Koulayev, Michael Manove, Andy Newman, Daniele Paserman and Johannes Schmieder.

I sincerely thank my wife, Ashley. I could not have completed this dissertation without her. Throughout my work, I have leaned on her constant support and limitless patience. I thank her for her continual encouragement, and her tolerance for the stacks of papers and scribbled on whiteboard that at least occasionally occupied our living room. I am thankful for her friendship and the privilege to share with her in this incredible adventure.

I owe so very much to my parents, Jay and Diane Young. I thank them for the years of guidance and unconditional support, and for the countless sacrifices they made of their time and resources to provide me with every opportunity. I have benefited from their strong work ethic and encouragement to embrace challenges. However, choosing to do everything the hardest way possible is probably of my own doing. I appreciate their encouragement to pursue my own interests, for fostering a passion for learning and a fascination with new thoughts.

I thank my grandparents for emphasizing the importance of education and the model they set with their dedication to serving their community and others. In particular, I thank Virgil Brown for his great mentoring. I wish to thank my sister, Lisa Kissel, for her friendship and support, as well as Mike and Elaine Brown, George and Cathy Young and Bill and Debbie Bethards for their encouragement and support.

I wish to thank those who have helped to set me on the path that brought me to this experience. They have sometimes been guides, mentors or a point in the right direction at a crucial juncture. I thank Peter Harder for the quality of my first undergraduate course in economics. I thank John Krieg for my second course, and many subsequent ones. I still have the exam on which he wrote, "Become an economist!" Derek Kilmer pointed me in a direction that led me to my path. I am grateful to Jon Orszag for the opportunity to pursue economics professionally, and in one of the most exciting environments in which I have ever worked. I thank Meg Guerin-Calvert and Mark Israel for investing in me during those years.

Finally, I wish to express my deep gratitude to the Department of Economics at Boston University for the opportunities it has provided me. I particularly want to thank my friends and colleagues, Amrit Amirapu, Michael Gechter, Debbie Goldschmidt and Anusha Nath. I value the conversations we've had, experiences we've shared and their impact on me as an economist and a person. I am also thankful to the department, the Institute for Economic Development and the Center for Finance, Law and Policy for funds to pursue my research, including work in India.

ESSAYS ON DEVELOPMENT ECONOMICS AND INDUSTRIAL ORGANIZATION

(Order No.)

NATHANIEL YOUNG

Boston University, Graduate School of Arts and Sciences, 2015

Major Professor: Dilip Mookherjee, Professor of Economics

ABSTRACT

This dissertation studies two disparate topics in development economics and industrial organization respectively: (a) the role of financial intermediation in promoting economic growth in developing countries and (b) the effects of learning on agents' search behavior.

The first essay investigates the effects of commercial banking on economic growth. The tendency of banks to locate in profitable areas experiencing higher growth typically complicates the identification of banking effects. I exploit a previously unstudied reform of bank branching policy in India between 2005-06 that led to a large expansion in private bank credit to financially underserved areas. Using iterations of a regression discontinuity design, I trace the exogenous expansion of banking services through time at the district level. I show this expansion produced positive effects in agriculture and manufacturing. I confirm greater gains in local GDP growth using remote sensing data to overcome the lack of official GDP statistics at the district level. These results offer evidence of a causal impact of financial system expansion on economic development.

The second essay examines how the geographic reach of a bank's network of branches can affect its ability to spread risks across spatially separated regions. I investigate the causal impact of the spatial expansion of Indian banks resulting from the bank branching policy reform on smoothing the consumption of households with respect to local weather and agricultural productivity shocks.

The third essay (coauthored with Sergei Koulayev) extends a model of sequential search for differentiated goods by relaxing the standard assumption of rational expectations. Agents are likely to refine their imperfect knowledge of product markets while searching. By introducing Bayesian learning into agents' beliefs, the model better replicates important aspects of search behavior. Using data from a popular internet hotel search site, we estimate lower median search costs in the model with Bayesian learning. Considering a counterfactual in which the first page of search results present the most popular hotel options, we estimate an increase in the number of successful searches.

Contents

List of Tables

List of Figures

Chapter 1

Formal Banking and Economic Growth: Evidence from a Regression Discontinuity Analysis in India

1.1 Introduction

Access to credit expands the choice sets of households and firms, allowing them to smooth consumption and investments across time. The early literature on access to financial markets establishes the association between strong financial systems and economic growth King and Levine (1993); Jayaratne and Strahan (1996); Rajan and Zingales (1998). The intuitive force behind this connection is that productive firms, particularly small ones, often lack the ability to self-finance and rely on external resources to achieve optimal levels of investment and growth. The extent to which such firms remain credit constrained can result in hindered economic growth at the aggregate level. An obvious policy prescription to facilitate growth would be to adopt policies that broaden access to financial markets. In the context of developing economies, this can often mean a literal spatial expansion of bank networks into unbranched or under branched markets.

The fundamental question is then whether policy induced branches actually affect firm and household access to credit. Though the answer may appear to be obvious and affirmative, a lower presence of the formal commercial banking system likely reflects underlying market characteristics. If banks are compelled under regulation to open branches in locations with high costs to doing business, extreme information asymmetries, or difficulties in aligning branch incentives with those of the bank, then new branches may fail to generate new bank business. Policies that do not address these underlying issues must instead focus on expanding branches in areas with a low banking presence explained by high fixed costs of entry. In either case, banks may simply be competing for market share with an informal lending sector, which can often be extensive in developing areas. If informal lenders adequately meet the needs of borrowers, formal credit expansion would be unlikely to produce additional growth.

Accounting for the concerns raised above, this paper examines the effect of expanding access to the formal banking system on economic growth. To gain deeper insight as to the potential channels generating growth that may lead to poverty alleviation, I investigate responses in agriculture and manufacturing to exogenous changes in the availability of formal credit. Examining a previously unstudied branching policy reform in India from 2005, I rely on institutional knowledge of the reform and India's banking environment to construct an empirical strategy and connect the timing of strong effects observed in the banking system to details dictating the reform implementation. The exogenous variation in access to formal banking generated by the policy reform facilitates the clear identification of the effects of banking on agriculture, manufacturing and growth. The seminal empirical paper on bank branching and development, Burgess and Pande (2005), exploits the timing of earlier reforms to show that a rapid expansion of the banking system in India during the 1977-1990 Social Banking Period corresponded with large reductions in the incidence of poverty. The analysis in the current paper identifies the potential deeper mechanisms underlying the effects of greater access to formal banking.

Identifying the effect of banking on specific channels of growth is generally difficult because banks tend to concentrate in profitable areas that are also likely to experience higher rates of economic growth. During times of policy intervention, banks may instead concentrate in poorer areas with slower growth. The bias from this endogeneity can vary widely, overestimating the impact of banks in the first circumstance and underestimating it in the second. Obtaining the necessary exogenous variation required to make causal inferences can be extremely challenging, particularly in development settings. Policy reforms, that may lead to natural experiments, if implemented at too fine of a geographic or demographic level may be impossible to evaluate for lack of granular enough data. Broader reforms may target areas receiving multiple interventions simultaneously making the effect of one particular mechanism inseparable from those of the others. Kochar (2011) studies the effect of banking on inequality in India during the Social Banking Period and raises just such a concern that the growth of subsidized credit distributed through the Integrated Rural Development Program (IRDP) correlates closely with branch expansion at the state level.[1]

[1] Kochar (2011)also discusses the timing and nature of branching policies from the Social Banking Period that are discussed in greater detail in an unpublished comment by Panagariya (2006) and a report by a

The identification strategy pursued in this paper constitutes a key contribution as it allows the use of district level data across all of India, while still separating the effect of the exogenous expansion in formal banking induced by the policy reform from other interventions introduced around the same time. I exploit the selection process of the reform that designated certain districts as being under banked and encouraged entry in these districts relative to those not receiving the status. The rule, based on district population per branch relative to the national average, generated an environment exploitable via regression discontinuity analysis using the national average as a cutoff and each district's population per branch as the running variable. Since the list of under banked districts remained essentially unchanged over the course of the policy, I am able to trace the policy effect on branching and credit, as well as agriculture and manufacturing, through time. I show the lack of a significant discontinuity in the pre-reform period, then a strong and accumulating expansion in private bank branches following the 2005 reform. I trace the effect through time by estimating the average treatment effect of the reform separately for each year from 2002 to 2012.

The timing of the policy response generates additional insights into bank behavior under the regulations. Due to policy details, banks were able to delay branch openings for a period following the policy implementation. Private bank branching in under banked districts remained low during that window, after which it steadily climbed. Meanwhile, the credit extended by private banks exhibited an immediate response consistent with the revelation of information that signaled a pending reform to branch licensing. The disparity in timing between branches and credit supply is consistent with the optimal strategy of profit maximizing banks that exhibit market power and anticipate future entry. Such banks may expand credit in anticipation of intensified future competition to lock in consumers who face positive switching costs. I provide a simple theoretical framework to outline the intuition for the incentives resulting from the confluence of the reform and India's banking

working group in the Reserve Bank of India (RBI) (RBI, 2009).

environment. I estimate the strongest responses to the reform from the private sector of India's banking system which is consistent with predictions based on the incentives generated around the cutoff. I cannot rule out, however, the possibility that the public sector also increased banking services due to the reform, but did so in districts away from the cutoff.

I draw on several different sources for the data in this analysis, including India's central bank the Reserve Bank of India (RBI), the Ministry of Agriculture, India's Annual Survey of Industries (ASI) and remote sensing data on the amount of light emitted at night and measures of rainfall. The detailed data from the RBI on bank branches and credit, from separate data sets, help provide a cross check for the two broad banking outcomes. Additionally, with credit reported annually at the district level with further disaggregation by bank group, population group and industrial sector, I am able to examine more detailed banking effects such as the expansion of credit to agricultural activities in rural and semi-urban areas following the reform. I combine separately reported data on district level crop production statistics and farm harvest prices from the Ministry of Agriculture to examine responses in agricultural productivity.

I conclude that the policy reform resulted in a significant expansion of banking services by the private sector in underserved areas. The cumulative effect of the reform is estimated as an average additional 10 private bank branches per district by the start of 2012. This constitutes approximately 50% of the sample average of operating private branches per district in districts around the threshold. I find evidence that private banks exploited the timing allowed in the policy to delay branch openings in under banked areas. In contrast, credit responded to early information regarding the reform consistent with banks racing to secure market shares. In 2006, the reform had already induced an average increase of 6,725 private bank accounts for under banked districts, approximately 52% of the sample average around the threshold. Growth in the credit extended for agricultural use in rural and semi-urban areas of under banked districts demonstrates that reform effects were not solely concentrated in high population areas. This is an important finding due to the popular concern that banks, particularly those from the private sector, exclude rural areas from service.

Increases to agricultural productivity that are consistent with the expansion of credit are observed. Agriculture constitutes a major employment activity in India, with over 56% of workers in 2001 engaged in agricultural endeavors. Further, policy makers placed particular importance on the availability of credit to rural and semi-urban agriculture leading up to the reform, suggesting early effects may concentrate in this sector. Positive responses in yield (output per hectare) and raw output are estimated for several important crops including cotton and wheat. These results are suggestive of a positive effect of banking on agricultural productivity, though a measure capturing responses across crops and incorporating harvest prices is preferred. Turning to an index of crop yield across several important revenue crops in India that weights by each crop's share of district revenue, I estimate an increase of 1,000 private bank credit accounts in a district raises average crop yield by 3%. This effect is a little less than half of the effect Jayachandran (2006) measures on crop yield from positive rainfall shocks, measured as district rainfall being above the 80th percentile of rainfall for that district. This result may reflect the ability of farmers to apply higher quality inputs purchased with credit, such as fertilizers or machinery. A redistribution of crop selection by high productivity farmers responding to loan availability may also contribute to this effect.

Effects are also observed in the amount of borrowing and production activities in manufacturing using data from the ASI. To analyze the ASI data that are available at the state level, I identify a set of treatment and control states based on the share of a state's population close to the threshold on one side or the other. I then perform a difference in differences analysis to estimate the effects of the reform. I estimate that enterprises in states with populations most affected by the reform experienced faster growth in the amount of loans they carried in the order of 23%. This result is consistent with the reform affecting the availability of credit to manufacturing. These enterprises also reported higher total investments, working capital and capital labor ratios following the reform. These responses are consistent with manufacturers being credit constrained in the pre-reform period and expanding their levels of capital given increased access to credit.

Finally, I confirm the aggregate effect on local GDP growth by showing that areas with expanding banking services experienced higher rates of growth in nighttime light intensity in the years following the reform. Henderson et al. (2012) established that so called "night-lights" provide a reliable proxy for economic growth under certain caveats when regularly reported data on traditional measures are unavailable, as is the case for district level GDP in India. Taking the estimated elasticity of nighttime light to GDP from Henderson et al (2012) of 0.3, the effect estimated in the current analysis implies that each additional private bank branch led to a 0.36% increase in local GDP. Overall, these findings offer strong causal evidence that the expansion of the financial system facilitates growth across productive sectors and encourages economic development.

1.1.1 Related Literature

These results are largely consistent with two analyses examining effects of branching in the United States. Dehejia and Lleras-Muney (2007) examine the effect of two forms of financial development in the United States from 1900-1940 on agricultural and manufacturing sector development. They find that while increased bank branching encouraged growth in these sectors, deposit insurance had negative effects. Krishnan et al. (2014) show that increased branching activity in the United States, following the Interstate Banking and Branching Act of 1994, led to greater efficiency gains by previously credit constrained manufacturers. Unlike the policy reform in the current analysis that directed banks toward targeted areas and resulted in policy driven branching, the mechanism of branch expansion in these other two analyses was legislation granting banks greater ability to branch, enabling banks already wishing to expand to enter new markets. The results in the current paper confirm the positive effects of branching on both agriculture and manufacturing. The three analyses differ in their consideration of manufacturing. Dehejia and Lleras-Muney (2007) focus on expanding labor demand in manufacturing; Krishnan et al. (2014) explain increased efficiency through the adoption of more productive projects, while the current analysis focuses on capital use, showing increases in investment and capital intensity in production. To the extent that firms remain capital constrained, lower aggregate TFP due to resource misallocation as argued for the case of India in Hsieh and Klenow (2009) may be partially attributed to lower financial access. The consistency between the broad effects is not surprising given that the expectation in the current context on the private sector banks is to compete and expand access to credit conditional on entry.

The differences that I observe between private and public sector banks, in their responses to the reform, likely reflects differences in the incentives and objectives generated under those respective ownership structures. La Porta et al. (2002) show that a higher incidence of government ownership in banking is correlated with slower growth looking across countries. In a series of joint and separate papers, Banerjee, Cole and Duflo examine the activity of banks from the public sector in India (Banerjee and Duflo, 2001; Banerjee et al., 2004; Banerjee and Duflo, 2014; Cole, 2009). They show evidence of under lending to productive firms, inertia in credit limits extended to firms and little difference in delivering development oriented lending resulting from government ownership. The main argument for these effects are misaligned incentives within banks, with loan officers facing few benefits from financing productive projects but punishments for loans that fail. De Quidt et al. (2013) demonstrate theoretically how market structure can greatly effect financial sector outcomes in the context of microfinance, comparing for-profit and non-profit lenders. The current analysis addresses these issues of bank ownership and incentives by analyzing private and public sector banking responses separately. The private sector, which is more likely to face profit maximizing objectives, shows a strong response to the reform around the cutoff. The public sector shows little evidence of a response near the cutoff, but this may reflect a different set of objectives that could concentrate their efforts in districts away from the cutoff, where my identification strategy does not apply. These findings highlight the importance of accounting for the institutional environment in crafting policy interventions.

In the next section I describe the important institutional aspects of India's banking system and the policy reforms to the branch licensing policies utilized for analysis. In section 3 I outline a simple theoretical framework to provide intuition for potential responses to the policy reform. In section 4 I review the regression discontinuity framework and describe how I translate its principles for analyzing the manufacturing sector with difference in differences. In section 5 I describe the data used in analysis. Then in section 6 I first establish a clear response in branching behavior to the policy reforms, then identify corresponding responses in aggregate private sector credit. I then examine responses in the agricultural sector followed by manufacturing. I then present results on overall growth using nightlights. Section 7 concludes.

1.2 Policy Reform and Institutional Background

1.2.1 Policy Reform

The Master Circular on Branching Authorisation Policy released September 8, 2005 implemented the policy reform on branch licensing utilized in this paper. The banking sector in India does not permit free entry of banking firms or branches. New bank licenses are granted infrequently by the Reserve Bank of India (RBI), India's central bank, through special campaigns with recent waves in the early 1990s and again in the early 2000s. Banks must also acquire licenses prior to opening all new branches, as well as receive permission to close or shift branches in most markets. Prior to the 2005 reform, banks applied for each of these changes on a case-by-case basis through the regional office of the RBI. No broad directive with regards to the composition of markets served by the bank, such as a requirement to open branches in rural areas, existed following the end of the Social Banking period in 1990.[2]

[2]The LEAD banking scheme was in operation during this time, however, by which one bank was assigned to each development block and made responsible for meeting agreed levels of branching and banking services. These banks were typically selected from the set of government owned banks. The service area approach

The reform in 2005 changed the regulatory environment in two fundamental ways. First, the reform effectively tied new branch licenses for highly sought markets to branch entry in markets designated as under banked. Specifically, banks were issued a set of criteria by which they would be judged during the review of proposed licenses. The "nature and scope of banking facilities provided by banks to common persons, particularly in under banked areas" would be considered when granting new licenses. In addition to offering "no-frills" bank accounts, meeting priority sector lending obligations and instituting a system for receiving and addressing customer complaints, banks were encouraged to open branches in "under banked districts and rural centres." The RBI provided a list of under banked districts with the circular. Though not stated explicitly, I will argue that a form of quota system operated requiring expansion in under banked districts for entry in rich markets. Second, the case-by-case application procedure for licenses was substituted with an Annual Branch Expansion Plan (ABEP) framework. Under the new system, each bank would prepare a set of proposed network changes (branch openings, closings and shifts) to be implemented over the next year. The plan would be submitted to the RBI for review, after which the bank management would meet with RBI officials to revise and finalize a set of permissions to be valid for the next year (Master Circular Branch Authorisation Policy, 2005).[3] The rule governing the assignment of under banked status was based on the district average persons per branch relative to the national population per branch for India (RBI Report 2009). The spatial implications of branch licensing from the reform around the national average cutoff provide the identifying variation exploited in this analysis and is discussed in detail in section 4.

Important differences exist between the above policy and those implemented under Social Banking. The degree of choice given to banks in selecting locations in which to open under the 2005 reform far exceeds that available during Social Banking. Unlike the 4:1

(SAA) also operated at this time, partitioning rural areas between banks for implementing development objectives.

[3]Permissions were valid for one year with the potential for extensions. Banks accomplishing 75% of their planned expansions could submit their next ABEP regardless of the lapsed time.

entitlement policy studied in Burgess and Pande (2005), that required intervention branches be opened strictly in unbanked markets, banks could choose among any markets within under banked districts to satisfy their obligation, allowing for the potential of increased direct competition between branches and banks. In stark contrast to the planned approach to district-wise branch expansion implemented in the 1980s (RBI Report, 2009; Kochar, 2011), banks under the current reform could choose between under banked districts for entry, as well as decide their own level of total entry, which affected their amount of entry in under banked branch districts.

Finally, the banking environment differed drastically in its composition and scope of business. The private sector, largely inert under social banking, expanded and gained vitality following the deregulations beginning in 1990 and infusion of "new private" banks. Government owned banks, consisting of the State Bank of India and its Associated Banks, the set of nationalized banks, and most regional rural banks (RRBs), have traditionally dominated the banking system in India. Following reforms and deregulation after a current account crisis in 1991, a sizable private sector developed, operating alongside government owned banks. The entering new private banks were heavily vetted and selected from many candidates during a period of open applications in 1993 and again in 2001. According to RBI documents, the purpose of these new banks was to foster competition and modernize the banking system. The new private banks broadly face the same regulation as the other scheduled commercial banks, though carry the additional mandate of maintaining at least 25% of their branch network in population centers with fewer than one hundred thousand people. The other policies they face, as well as their requirements to the Priority Sector lending scheme, are identical to those on the SBI and Nationalised banks. RRBs and foreign banks face tailored regulations, including those pertaining to branching requirements.

1.2.2 Policy Details and Timing

While the reform became official in September 2005, events leading up to its release likely provided signals as to its impending introduction. In a speech from December 2002, the director of the RBI pointed to the high share of bank investments in government securities, 39% relative to the regulatory minimum of 25%, encouraging banks to expand their commercial lending particularly in small manufacturing and agriculture (Mohan, 2002). The following November, the Vyas Committee was commissioned to investigate the flow of capital to agricultural activities. They met with several commercial banks during their investigation. In April 2004, they released an interim report followed by the final report in June, suggesting revisions to the service area approach (SAA) and encouraging greater lending by private and public sector banks. The report included a map identifying areas underserved by the formal banking sector, some of those identified as places where the "branch network of commercial bank[s] [is] below the national average (Vyas Committee Report, 2004)." The SAA program was subsequently discontinued, allowing all banks to freely apply for entry and operate in rural areas. The official list of under banked districts released in 2005 almost exactly matches selection based on district average population per branch relative to the national average, consistent with the language in the report. Thus, aspects of the Vyas Committee report could have provided solid signals to banks of the forthcoming reform.

The list of under banked districts initially released in 2005 was reissued in 2006 adding a small set of districts that satisfied the under banked requirement in both years but were left off of the 2005 list. Afterward, the list was reissued each year unchanged until 2010.[4] After 2010, certain states were made ineligible for under banked status, reducing the number of

[4]Starting in 2008, certain centers within under banked districts were made ineligible to count toward a bank's serving of common persons. Specifically, centers within the municipal limits of state capitols, district headquarters and metropolitan centers were deemed ineligible. Further, centers within 100 km of Mumbai, New Delhi, Kolkata and Chennai, and 50 km of state capitols were ineligible. Exceptions were made for the state of Jammu and Kashmir, and the seven North Eastern states, Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland and Tripura.

districts considered as "under banked districts of under banked states," but not introducing any new districts to under banked status. Although additional reforms altered the incentives for branch expansion both within and outside under banked districts, given the lagged nature of branch openings to license issuance, I would expect to and do find lasting effects through 2012. In section 4, I discuss the algorithm used for assigning under banked status to districts in detail and how I exploit it following a regression discontinuity design strategy to identify the effect of exogenous expansion in formal banking on real economic outcomes.

Although the reform became effective immediately upon its release, banks were essentially allowed a year long grace period to construct their first ABEP, with an implicit deadline for September 2006. Several banks, many of them from the private sector, waited close to the full year to submit their ABEP, during which time they were able to receive licenses in a disaggregated manner. The histogram of branch license dates for a large private sector bank is shown in figure 1.8.[5] Although annual branch expansion plans may not be observed directly, the large spikes in branch licenses set approximately a year apart are consistent with ABEPs. The figure shows the licenses from the first likely ABEP for this bank were granted in July 2006, roughly one year after the reform implementation. Similar patterns are identified for many private sector banks. Additionally, the licenses from ABEPs remain valid for a year, meaning that banks could effectively postpone the effect of the reform for nearly two years if preferable to quick entry. The optimal timing for entry from the perspective of the banks will depend on strategy and the underlying profitability of the locations. Early entry may allow banks to secure market shares, though they could delay costly entry into low profitable areas by waiting. The empirical evidence suggests most private sector banks chose to delay entry in locations of induced entry.

Finally, the shifting and closing of branches, particularly in under banked districts, was heavily regulated. Branches were not allowed to shift outside otherwise unbanked centers. Given the source location was served by another commercial bank branch (other

[5]Known acquisitions of branches from other banks have been excluded for the histogram analysis.

than a RRB), a branch could only shift to centers in the same or lower population group classification, and in the case of branches in under banked districts, could only shift to centers within under banked districts. Little branch closure is observed in the data, though mergers and acquisitions of banks occur during which most branches are "closed" and "reopened" under the acquiring branch, with some branches converted to satellite offices and fewer still permanently closed.

1.2.3 Policy Reform Discussion

Incentives The 2005 branch licensing policy reform purposefully created new incentives for scheduled commercial banks to open in centers conditional on their district's under banked status. Licenses for branches in high profit potential centers in banked districts were used to leverage bank entry into under banked districts. This mechanism works most effectively during periods of high demand for bank branches in "rich" areas, as was presumably the case experienced in India during its time of high economic growth beginning in 2003 and continuing through the decade.

The branching policies and reform placed no requirements on the amount of banking required to occur at each branch. There are staffing requirements for branches, as well as minimal days and hours of operation. Banks must also offer "no-frills" accounts that carry limited fees and low minimal balances to prevent the exclusion of poor customers. Despite these requirements, though, banks could maintain staffed branches that simply minimized costs by not reviewing or approving any loan applications, not pursue new customers, and only accept deposits.

Banks are also required to meet Priority Sector lending ratios. Banks must maintain 40% of their outstanding credit in loans to the priority sector. However, the requirement must only be met at the bank level, meaning some branches may carry heavy amounts of priority sector loans while others lend nothing to the priority sector. In 2007, new guidelines were adopted for the priority sector, reducing the set of loan categories eligible for priority status.[6] The reformed guidelines concentrate lending in direct and indirect agricultural endeavors and limited the amount going to microfinance institutions and other indirect modes of lending. The priority sector reforms apply uniformly at the national level. Banks failing to meet their 40% requirement must make up the difference with loans to the NABARD RIDF fund at deterrent rates. Banks typically come very close to meeting the requirement, overshooting slightly in some years and falling short in others.

1.3 Theoretical Framework

This section articulates a simple theoretical framework to provide intuition for the effects of the reform and the heterogeneity of responses across districts and bank groups. The theoretical framework demonstrates how the 2005 policy reform may have incentivised higher rates of entry in under banked districts and increased lending without addressing the underlying profitability conditions of those districts. Further, it rationalizes an expansion of credit in under banked districts following the policy announcement, prior to a significant increase of branches.

The framework begins from a standard characterization of financial intermediation with adverse selection of borrowers common to credit markets in developing economies. Consider a single market with two periods and two types of borrowers, safe and risky. In the first period, a policy reform that will encourage entry in a (potentially unknown) set of markets beginning in the second period is announced. In the second period the reform is in effect. As in Stiglitz and Weiss (1981), each borrower has a potential project that requires a loan (normalized to size one for all borrowers) and yields the same expected return across borrowers. The borrower is assumed to have the same potential project in each period. Assume that the return from a failed project is zero, so that $P_s(R_s^A)R_s^A = P_r(R_r^A)R_r^A$, where R_i^A is the return from a successful (denoted A) project for type $i \in \{safe\,(s), risky\,(r)\}$ and$P_i(R_i^A)$ is the probability of success for type i. Thus, safe types have projects with lower

[6]The reforms to the composition of the priority sector studied in Banerjee and Duflo (2014) occurred in 1998 and 2000, prior to my analysis.

returns conditional on success but succeed with greater probability $P_s(R_s^A) > P_r(R_r^A)$. If banks operate in the market, they can offer a standard debt contract with fixed repayment. Assume $R_i^A > (1 + r_i) > 0$ and that borrowers face limited liability, such that when a project is successful the borrower pays back the principle on the loan plus interest at rate r_i, but that in case of failure no payment is made and both borrower and bank receive zero. Borrowers face an outside option that provides utility equal to μ. Both borrowers and banks discount the future at rate δ and are risk neutral. While borrowers know their own type, banks only know the distribution of types and the parameters defining the projects. Banks prefer to lend to the safe types due to limited liability but cannot distinguish between types in the general framework. Depending on the set of parameters just described and the share of safe and risky types in the population, banks may choose to ration credit in response to adverse selection, or the market may collapse entirely (Stiglitz and Weiss, 1981).

To capture the dynamic effect of the policy reform, consider the two following modifications: 1) Banks possess a screening technology that reveals a potential borrower's type with certainty and costs amount s. 2) There exists a downward sloping demand curve among safe types. A wide set of assumptions can satisfy this condition, for example, if personal costs of marketing the successful project differs between borrowers then demand for loans will decrease in r_s. To simplify the analysis, assume parameters are such that banks always choose to screen borrowers and never find it profitable to lend to the risky types. Further, the banks pass the cost of screening on to the borrower. As long as borrowers must repay the full amount of the loan conditional on a successful project, and borrowers cannot accept contracts with the potential for negative consumption, the expected default rate from safe types will remain unaffected and banks will know the demand conditional on the interest rate offered with certainty.[7] This assumption greatly simplifies the game as it allows the borrower's decision process to be considered separately for each period since current nega-

[7]A contract with potential negative consumption would arise when limited liability protects the borrower against a failed project, but not from a successful project for a borrower whose high marketing costs leaves them less from the project than the fixed payment owed to the bank.

tive expected returns cannot be offset by more favorable expected lending conditions in the future.

Assume banks are symmetric and profit maximizers, each facing an exogenous marginal cost of funds, including administrative costs from lending, equal to $(1 + \rho)$, and cannot discriminate in the interest rate it offers to repeat versus first time borrowers. Since banks observe the parameters on the population defining the distribution of safe types, they know the slope of the demand curve, though do not know any particular borrower's value of the loan. Without the threat of entry, a monopolist serving the market in the first period maximizes profits by serving the same set of borrowers in each period, increasing the interest rate in the second period to capture the additional surplus the borrowers receive from not paying the screening cost again (a sketch of the proof is given in the appendix). Knowing this, the monopolist may work backwards from the second period to determine the profit maximizing interest rates in each period. In contrast, when two banks serve a market, they compete in prices in both periods. If both enter the market in the same period, then each offers the zero profit interest rate and split the market.

However, if one bank acts as an incumbent, then it may choose to alter its behavior when anticipating the potential of entry. The screening cost, which may be construed as the cost to the borrower of filling out paper work or the effort of establishing a relationship with a loan officer, operates as a switching cost for the borrower.[8] Borrowers will go to whichever bank results in them keeping the highest expected return from their project. For first time borrowers this is simply the bank offering the lowest interest rate. Repeat borrowers must compare their expected payoff from the incumbent's 2nd period interest rate to that of the entrant plus the screening fee required to switch. The resulting equilibrium is intuitive: in the second period, under cutting leads the entrant to offer the zero profit interest rate and the incumbent offers an interest rate making its set of first period borrowers indifferent between switching to the entrant and staying. Since the set of first period borrowers is

[8]Klemperer (1987) lays the groundwork for considering the effect of switching costs.

entirely determined by the first period interest rate, the second period interest rate is a function of the first period interest rate and the screening cost. Knowing this, the incumbent chooses the first period interest rate that maximizes profits over both periods. The threat of entry will result in the monopolist offering lower first period interest rates to secure a larger base of customers from which to earn positive profits in the second period. The set of parameters will determine how willing the incumbent is to trade off first period profits for those in the second period. The entrant will serve the remainder of the market that demands loans at the zero profit condition. Thus, credit will initially expand with the announcement of the policy reform and again upon realized entry.

1.3.1 Entry

The effects on entry must be primarily driven through changes to the structure of fixed costs of entry as the reform did not otherwise target local market conditions. Consider multiple markets described by the framework above. Markets are differentiated by their set of parameters already discussed plus overall market size. Suppose banks each draw market specific fixed costs of entry for every market. Abstracting from the strategic considerations of entry, assume banks act myopically such that they expect to act as a monopolist if entering a market unbanked in the first period or as a duopolist when entering banked markets. Under these assumptions, expected profits for each market is known to a bank and entry will occur for all markets j satisfying $E[\pi_B^j] - F_j > 0$. Markets with low profit potential or high fixed entry costs will fail to attract banks.

Consider a rule that ties permission for entry in some high profit potential markets to entry in lower profit ones. Banks facing binding constraints will now open into markets where $E[\pi_B^{UB1}] - F_{UB1} < 0$ if these losses may be offset by the profit gains from the rich market, $E[\pi_B^j] + E[\pi_B^{UB1}] - F_j - F_{UB1} > 0$. This condition will be more easily satisfied in policy eligible districts with higher expected profits that faced unfortunately high fixed entry costs. Once entered, however, these markets may produce high levels of banking activity. In contrast, the set of markets originally served without the reform may contract if the lowest profit earning locations cannot offset the losses from policy eligible markets. Finally, the joint positive profits will be hardest to satisfy for policy eligible districts that face the lowest profit potential and highest fixed costs of entry. The reform will be unlikely to produce positive banking results for such markets.

1.3.2 Predictions

The above framework suggests three main predictions of bank responses to the policy reform. First, the amount of credit will expand in districts where increased entry under the reform is expected to occur. This will result from banks attempting to secure market shares in these districts so as to lock borrowers in through switching costs prior to the intensive entry under the reform. Second, branch entry will increase in under banked districts with high profit potential that was offset by high fixed costs to entry which become subsidized due to the reform. Further, entry may be most profitable in locations where banks open as an entrant, with lower fixed costs making up for stronger competition for borrowers. Thus, both entry as monopolists and as competitors is possible. Finally, this framework characterizes behavior under profit maximization. Banks following other objective functions, as public sector banks may do, would be less likely to generate these responses. Comparing the behavior of private sector and public sector banks will provide a qualitative test of predictions from the framework.

1.4 Empirical Methodology

Identifying the effect of bank branching on banking and real economic outcomes can be frustrated by classic endogeneity concerns outlined in previous work Burgess and Pande (2005), in which selection bias can overpower estimates, even changing their signs. The unique policy aspects of the 2005 branching reform create an environment facilitating the identification of banking effects on agricultural and industrial outcomes. I am able to circumvent endogeneity concerns and separately identify the banking effects from other simultaneously operating reforms by employing a regression discontinuity design that yields transparent estimates and identification founded on assumptions that are at least partially testable. First I identify and quantify the expansion of banking services in response to the policy. Next I focus on the effects of banking in agriculture, which appears to be an initial motivation for the reform and the largest employment activity in India. Then I turn to the effect of banking on manufacturing enterprises, which appeared to gain from the realized expansion of bank branches. After establishing a response in these two areas, I provide evidence of a positive effect on overall growth using light emitted at night as a proxy.

1.4.1 Regression Discontinuity

The method employed by the RBI for identifying districts as under banked in the 2005 branching policy reform, based on simple district and national averages of population per branch, yields a clear quasi-natural experiment exploitable by regression discontinuity techniques. Under banked districts were identified using two inputs. First, the national population of India, taken from the Population Census conducted in 2001, was divided by the total number of scheduled commercial bank branches operating in the country in 2005-2006 to obtain a "national average of population per branch." Then an analogous value was calculated for each district and compared to this national average. Those districts with a calculated value higher than the national value were designated under banked. Figure 1.4 shows district under banked status from the 2006 list of under banked districts plotted against district population per branch. According to the rule, districts to the right of the cutoff should be assigned to under banked status, as is broadly confirmed in the graph.[9] A map of the districts in India with their corresponding district averages is presented in the upper panel of figure 1.5.[10]

[9]Six districts do not follow the assignment rule, with four of them remaining in the sample used in estimation (see the section on constructing the forcing variable in the Data Appendix for details).

[10]The districts with greater deficits of branches per person, denoted by darker colors, matches closely with the areas identified as being more broadly under served by the map from the Vyas Committee issued in

The above algorithm induces a cutoff at the value of the national average, treating district population per branch as the "forcing variable." The policy generates an arbitrary difference in districts falling on the "under banked" side of the cutoff, which offer an additional value to banks opening branches within their borders: such openings count toward their requirement for "serving common persons" to gain permissions for branches in rich markets. Districts falling on the other side of the threshold do not offer this benefit, though being otherwise similar, which will be tested formally. Thus, the policy effects the probability that the districts will receive additional branches through its manipulation of bank incentives. This estimation strategy will be valid if the distribution of potential outcomes is continuous at the cutoff(Lee, 2008). A lack of perfect manipulation of the running variable so as to change a district's treatment status, and the continuity of other factors that may affect the outcomes of interest with respect to district population per branch near the cutoff will suggest this assumption is satisfied. I verify that both of these stipulations hold below.

Figure 1.6 presents visual results from the McCrary test for manipulation of the running variable around the threshold McCrary (2008). The distribution of districts along the running variable is shown to be smooth around the threshold. The discontinuity estimate in the log difference in height is 6.6 with a standard error of 22, thus I fail to reject the null hypothesis of continuity. The figure also highlights another ideal trait of this environment; the cutoff is located near the peak of the density, meaning most districts fall close to the cutoff, suggesting the generalization of the effect for most districts may be reasonable. The lack of manipulation around the cutoff, beyond passing the McCrary test, is extremely defensible on intuitive grounds. Even if banks and districts were able to perfectly anticipate the criteria for assigning under banked status, their ability to manipulate assignment would be limited. The population level in the current equation was taken in 2001, four years prior to the policy. Thus, agents attempting to influence district status could only do so through altering the number of operating branches within district boundaries, which results from

2004.

the collective branching decisions of all banks and conditional on RBI permissions, making manipulability extremely unlikely.

Figure 1.7 presents a series of plots of district baseline characteristics, with dots reporting local averages for districts falling within 200 persons per branch non-overlapping bins. A local linear regression of the data is shown with flexible slope on either side of the cutoff. While the figures constitute a visual RD testing for continuity at the cutoff centered at zero, they also summarize broader trends in branching at the time of the policy reform. Districts left of the cutoff enjoyed more branches per person by definition. These districts also tended to be places with higher populations living in large cities, exhibited higher literacy rates, had lower populations of scheduled caste and tribe persons and had a lower percentage of main workers engaged in agriculture. Each of these characteristics appears to be smooth at the cutoff, suggesting proper randomization of districts around the cutoff. The continuity is tested formally by performing RD analysis with the baseline characteristics as the dependent variable. The tests fail to reject the null hypothesis of continuity at the threshold, with reduced form results presented in table 1.1.

1.4.1.1 Technical Details of RD

The identification of local average treatment effects through regression discontinuity analysis is now well established in the literature (Black, 1999; Angrist and Lavy, 1999; Van der Klaauw, 2002; Lee et al., 2004), with the theoretical work on identification in Hahn et al. (2001) and the origins of the method in Thistlethwaite and Campbell (1960). To reduce bias from including observations far away from the cutoff where the identification does not hold, I use local linear regressions, dropping observations outside a set bandwidth of the cutoff (Hahn et al., 2001; Lee and Lemieux, 2010). I restrict all analysis to local linear and local 2nd degree polynomial regressions as recommended in (Gelman and Imbens, 2014). I set the bandwidth at 3.5 thousand persons per branch for all regressions, which falls within the range of optimal bandwidths selected for individual years by the Imbens and

Kalyanaraman (2011) method.[11] I fix the bandwidth to provide transparency for tracing the evolution of the policy effect across years, as this fixes the set of included districts across regressions. In Figure 1.5 the lower left map indicates districts included in the local linear regressions. The districts are spread out geographically across most of the country, with under banked districts typically not far from banked counterparts. The map in the lower right panel identifies districts close to the cutoff on either side. Again, districts are distributed geographically and tend not to bunch by under banked status.

For each year, I first estimate the local linear regression of the reduced form equation,

$$y_i \quad = \alpha + D_i\tau + f(PopPerBranch - Cutoff) + \delta X_i + \epsilon_i \tag{1.1}$$

using a uniform kernel. y_i denotes a banking or economic outcome of interest in district i, such as the number of operating bank branches or crop yield. $D_i = 1[PopPerBranch_i - Cutoff \geq 0]$ is an indicator for satisfying the rule for assignment to under banked status, $PopPerBranch_i$ is the population per branch for district i, $f(\cdot)$ is a flexible functional form, X_i is a set of controls, τ is the coefficient of interest measuring the discontinuity at the threshold, and ϵ_i is an idiosyncratic error. In all regressions, I include the pre-random assignment value of the dependent variable from 2001 to improve precision and reduce sampling variability (Imbens and Lemieux, 2008; Lee and Lemieux, 2010). In addition, I include the 2001 district population and its square to further improve precision. The described method constitutes the reduced form estimate, with the probability of under banked status instrumenting for actual assignment. These values are reported graphically.

I report the "fuzzy RD" results implementing the regression discontinuity using Calonico, Cattaneo and Titiunik's "rdrobust" package with a triangular kernel. I use the fuzzy RD with banking outcomes because the rule I use to assign under banked status does not per-

[11] Results are robust to different bandwidth selections, and 2nd degree polynomials typically perform better with wider bandwidths than linear specifications as in the example from Lee and Lemieux (2010).

fectly match the list.[12] When considering agricultural and other economic outcomes, their most interesting relationship is to realized banking behavior, rather than district assignment to under banked status. For these outcomes, in addition to the reduced form that presents the general effects of receiving under banked status, I estimate the effects with the fuzzy RD instrumenting for banking outcomes. That estimate will inform the effect of the specific banking outcome on the outcome of interest. However, that effect should be treated with caution as reform status will effect several dimensions of banking outcomes. To implement the fuzzy RD analysis I first "residualize" the data, regressing y_i on the set of controls X_i from equation 1.4.1.1, then estimating equation 1.4.1.1 replacing the left hand side variable with the residuals obtained from the first regression and dropping the controls from the specification (Lee and Lemieux, 2010). Conventional estimates of the RD are reported, as are bias-corrected estimates and the robust standard errors from Calonico et al. (2014). I will focus on the conventional estimates in discussing results.

1.4.1.2 Dynamic Strategy

The identification of the policy effect on banking outcomes is bolstered by the ability to regularly estimate the effect of the reform through time, both before and following its implementation. In the pre-reform period, no discontinuity should exist at the cutoff. In the post-reform period, the effect of the policy should be expected to grow according to the timing set in place by the rules of the reform and its revelation. To demonstrate the timing of the reform effects, I estimate equation (1.4.1.1) separately by year for banking outcomes, agricultural outcomes and outcomes measured through remote sensing including rainfall and nighttime light emitted into space. Given that the set of under banked districts remained essentially unchanged in the reform period considered, this captures the short and medium term effects as they emerge.

[12] I fail to match 6 out of 572 districts to their realized under banked status from the 2006 list. See the data appendix for details.

1.4.2 Effects observed in Manufacturing

To examine the effect of increased financial access on the manufacturing sector, I use ASI data available at the state level. The level of aggregation prevents conducting the regression discontinuity just described. Instead, I follow a difference in differences approach, utilizing the institutional knowledge of the reforms to construct sets of treatment and control states.

I select the set of "under banked treatment states" in the following way. Using population census data at the district level, I construct the shares of state population in under banked districts. For the population of each state in under banked districts, I calculate the share of that population belonging to districts falling within a close bandwidth of the national average of population per branch, generally within 4 thousand persons per branch. Those states with large shares of their population in under banked districts close to the threshold are selected as the treatment group. I then construct a control group using a comparable procedure from districts with banked status. "Banked States" include Haryana, Uttarakhand, Punjab, Mizoram, Daman and Dimiu, Karnataka, Puducherry, and "Under Banked States" include Rajasthan, Tripura, Jharkhand, Orissa, Dadra and Nagar Haveli.

For each treatment and control group pairing, I estimate the following,

$$\begin{aligned} y_{it} = \quad & \alpha + \xi post06_t * treat_i + \varphi post06_t + \psi treat_i \\ & + \beta_1 year_t * state_i + \beta_2 year_t + \beta_3 state_i + \beta_4 X_{it} + \omega_{it} \end{aligned} \tag{1.2}$$

where $post06_t$ indicates financial years 2006 and later, $treat_i$ indicates that the state belongs to the treatment group, and the remaining terms indicate controls for state fixed effects and state specific time trends, as well as a matrix of additional controls in X_{it} with an idiosyncratic error ω_{it}. The coefficient of interest will be on the interaction term $post06_t * treat_i$, which will give the difference of within-state differences between the states receiving under banked status and those not. In addition to controlling for post 2006 and treated state individual effects, the regressions include the logged number of manufacturing units in the firm and the logged number of employees in the enterprise to control for enterprise size. Plant age and its square are also included as controls as these may influence the firms' access to credit and capital markets. Although this identification strategy is not as ideal as the RD, the careful selection of the treatment and control states should help in eliminating potential threats and I will take the estimate as suggestive of the effect from the policy reform on manufacturing.

1.5 Data

The primary data on banking are from data sets maintained by the RBI. The Master Office File (MOF) provides a detailed record of bank branch locations and characteristics, from which detailed branch network information by bank may be constructed. I have also matched most branches to approximate geocoded locations based on postal codes (PIN) and center names. In the present analysis, the MOF is used to construct the running variable and national average based on population per branch at the district and national levels respectively. The Basic Statistical Returns 1, 2 and 7 provide time series data on credit and deposits at various levels of aggregation. The empirical methods and analysis pursued in this work is greatly determined by the level of data availability. Although branch location data are available in detail through time by bank, much of the credit and deposits data are only available annually as aggregates to bank group level by district. Thus, although it is possible to observe which banks are operating within a district, assigning a certain number of accounts or amount of credit from any particular branch is impossible. Fortunately, the policy reform applied at the district level, allowing analysis directly at the level of the reform. Utilizing the time dimension further helps to disentangle some effects of the reform from changes to bank group classifications.

The availability of credit limits, amounts and accounts by the intended geographic utilization of loans constitutes a strength of the data used in this analysis. The use of Call Reports from banks do not typically allow for this level of geographic precision of where loans are actually directed. This feature strengthens the arguments that loans reported in certain areas are not financing projects in neighboring districts.

To conduct the analysis on agriculture, I develop a new data set by processing and combining separate annually available data from the Ministry of Agriculture, Directorate of Economics and Statistics on crop production statistics and crop farm harvest prices. By matching district production levels to farm harvest prices by crop, I am able to construct an index of crop yields similar to that in Jayachandran (2006) for crop years 2002 - 2008. The use of an index circumvents certain concerns arising from differences in crop suitability across districts.

Data on manufacturing enterprises are from the Annual Survey of Industries, reported annually for registered firms. Measures from enterprises with fewer than 100 employees are taken from a 20% sample of firms representative at the state level. The ASI data used in this analysis does not report the district of the enterprise. As described in the empirical strategy section above, I adjust for the level of the data being broader than the level of the reform so as to best capture the spirit of the RD design.

District level data on several measures of interest, local GDP for example, are unavailable or available only sporadically. To overcome the lack of traditional measures, I consider data recorded from remote sensing on rainfall and the amount of light emitted at night from the TRMM satellite and DMSP-OLS Nighttime Lights Time Series, respectively. The nighttime light data are used to proxy for changes in local GDP, as prescribed in Henderson et al. (2012). See the Data Appendix for greater detail on all data used in the analysis.

1.6 Results

1.6.1 Banking

1.6.1.1 Bank Branches

The analysis focuses attention on the response from banks in the private sector. The notion that these banks introduce a new banking technology and the rapid expansion of their branch networks during this period makes them particularly likely to drive innovation and a transformation of the banking environment in affected districts. As profit maximizers, the theoretical framework suggests they are also the most likely to respond strongly to the reform around the cutoff. Responses from the public sector will be noted to provide contrast. To motivate the primary set of empirical results, I first consider a visual example for two years. Figure 1.9 presents the standard visual RD for operating private bank branches for the pre-reform year 2000 and the post-reform year 2012. The y-axis denotes the number of operating private bank branches per district, with dots reporting the local averages of districts falling within 200 persons per branch non-overlapping bins. The horizontal axis is the forcing variable of district population per branch centered on the national average and scaled to thousands of persons per branch. Considering the figure from year 2000, districts do not appear to vary systematically in their number of bank branches prior to the reform. In the post reform year, under banked districts show higher numbers of operating branches relative to banked branches just on the other side of the cutoff. The discontinuity of the number of branches estimated at the cutoff from either side yields the local average treatment effect of the reform on private branches. Next, I make the analysis more precise by presenting the annual results from estimating equation 1.4.1.1 with operating private branches as the dependent variable.

The ability to observe the number of branches across time, and the fact that the list of under banked districts did not change yearly, allows the effect of the reform to be identified not only by spatial variation between districts, but through time as the reform became implemented and branches were able to accumulate. The right panel of figure 1.10 plots the intercept points at the cutoff from annual local linear regressions from the banked and under banked sides of operating private sector branches in a district. Districts maintain the same value of the forcing variable across years so the set of districts remains unchanged.[13] The red

[13]New districts since 2001 that claimed territory from more than one source district are dropped along with the source districts. In addition, Thane and Pune districts in Maharashtra are dropped, as is Varanasi district in Uttar Pradesh after 2002. See the Data Appendix for details.

dashed line provides the estimated intercept from approaching the threshold along the under banked side as in the classic RD graphical representation. The solid blue line reports the corresponding intercept approaching from the banked side. The vertical distance between the two, reported for each year, corresponds to the discontinuity at the cutoff estimated as τ in equation 1.4.1.1. A vertical red line between the two points indicates a positive discontinuity, with under banked districts exhibiting a higher value at the threshold than banked districts, with significance at least at the 10% level.[14] These figures not only present the average treatment effect, but place the level of the intercepts vertically so that overall growth and decline may be easily recognized.

The figure identifies important policy aspects. In the years leading up to the reform, there is little difference in the estimated number of branches from the banked and under banked districts at the cutoff. This in itself acts as a partial validation test of the randomization of districts around the cutoff. A strong response to the policy does not occur until after 2006, which was a likely possibility given the timing of the reform. Still, the small increase in the positive discontinuity in 2005 and 2006 is not inconsistent with some banks working to establish market share in under banked districts. The strongest effects in branches occur beginning in 2008 and are estimated precisely at the 5% and 1% confidence levels, which is consistent with banks waiting until mid 2006 to submit their first ABEP and opening their branches just before their licenses expire in mid 2007. Estimates are presented in table 1.4. The steadily growing discontinuity is consistent with a response from private banks to the branching policy.

In the figure on the left, I report the estimated effect on operating and granted licenses. The most important feature from this graph is the first statistically significant positive effect on licenses measured one year earlier than branches on January 1st, 2007. Turning to the estimation results from licenses and branches in table 1.4, the effect from licenses precedes a similar response in branches beginning in 2007 through 2010. The policy starting in 2010

[14]Thanks to Johannes Schmieder for help in clearly displaying the dynamic nature of the effect graphically. Note that these figures rely on estimation using a uniform kernel.

was amended such that banks could open in lower population centers without a prior license, resulting in licenses for such openings being issued on the day of branch entry, despite its presumed reporting in ABEPs.

The combined timing of the licenses and operating branches, as well as the pre-reform and post-reform pattern demonstrates the exogenous differential change in branch reach in districts belonging to under banked districts near the cutoff relative to the otherwise similar districts on the banked side. The cumulative average effect of the policy in 2012 is estimated at approximately 10 more private sector branches in under banked districts at the cutoff relative to the banked districts. The effect is a little less than 50% of the sample mean reported in the table for 2012 at 20 private sector branches in districts around the cutoff. The size of the private sector presence increased for the sample overall in this time from an average of 10 branches per district in 2006 to 20 in 2012.

1.6.1.2 Credit

The 2005 policy reform on branching permissions directly cites opening branches in under banked districts as a condition affecting total permissions to a bank. However, the other terms mentioned, offering no-frills accounts and meeting priority sector requirements, apply at the bank level rather than by district. Thus, there is little direct pressure from the reform on bank credit and deposit behavior, particularly around the threshold. Recall, however, that the theoretical framework predicted that banks would expand credit in under banked areas in anticipation of future competition.

Figure 1.11 shows the annual discontinuities in total district credit from private banks, analogous to the figure presented for operating branches discussed above. Similar to the early pre-reform years in private bank branches, the number of credit accounts in thousands shown in the left panel of the figure displays little difference between the banked and underbanked districts at the cutoff. However, consistent with the timing of the Vyas Committee commissioning and report, the number of credit accounts began increasing in under banked districts in 2004 and 2005. Underlying this change is also a change in the composition of banks in these districts, with fast growing branches opening in these districts as more inert banks were acquired by nationalised banks. This behavior is consistent with aggressively growing banks acting preemptively on the expectation of reforms by expanding in areas likely to be more heavily contested in the future. Turning to the estimates in table 1.5, the response from this preemption is estimated at 6,725 additional credit accounts in the under banked districts at the cutoff, which is 52% of the sample mean for districts around the cutoff. The discontinuity in accounts continues to grow over the next few years and is estimated precisely for all years except 2009 and 2011. The slight retraction of the discontinuity in 2008 may be explained by the exit of a private bank through acquisition by the public sector in 2007.[15] The decline in the discontinuity in 2009 may reflect the tightening of restrictions regarding specific cities eligible as under banked within districts based on their proximity to major metropolitan areas or being metropolitan themselves. Unlike the branches data, credit cannot be broken out by bank within a district to form a clearer picture as to the exact channels driving the aggregate responses. The last two years again show increased expansion in credit accounts consistent with the growth in branches in these years.

The results from outstanding credit amounts in millions of rupees show qualitatively consistent results. The amounts data are measured with less precision, which may result from many large investment projects being lumpy in nature, leading annual district levels to fluctuate more than the number of accounts. [16]

As noted by RBI Deputy Governor Mohan in a 2006 speech regarding financial inclusion, the expansion of retail credit after 2003 accounted for a major source of increased lending (Mohan, 2006). Breaking credit out by personal loans, figure 1.12, confirms that the response in the growth of personal loans in under banked districts near the cutoff rela-

[15]Bharat Overseas Bank was acquired by Indian Overseas Bank that already held a 30% interest in the bank.

[16]The large dip in credit to banked districts in 2008 appears to be driven by outliers, as changes in districts affected by the above mentioned merger in the previous year do not show strong responses in credit amounts.

tive to banked districts was significant. The initial jump in personal loan accounts in 2004 corresponds with the changing composition of banks as aggressive private banks slowly expanded their branch presence in under banked districts. Also at this time, the interest rates on consumption loans were liberalized, allowing interest rates to dip below the bank's self reported cost of funds plus profit margin. Personal loan amounts largely mirror the expansion of accounts, though the tightening after 2008 may correspond to a change in priority sector lending requirements making the requirement more stringent.

An implication of the theoretical framework behind the hypothesis of a preemptive response in credit by profit maximizers is that public sector banks, which follow less clear objective functions, are unlikely to show the same pre-reform response as private sector banks. Figure 1.13 and the corresponding table 1.7 confirm a lack of response prior to the policy implementation around the cutoff, as well as a muted response during the reform years as well. These results are consistent with the incentives generated by the reform operating most strongly on private sector banks aiming to grow in reform years.

1.6.2 Agriculture

Agriculture constitutes the primary economic activity for the majority of Indians. The 2001 Population Census reports that over 56% of India's workers were engaged in agricultural or related activities at the time of the census which, due to the exclusion of marginal workers, likely provides a lower bound. Given that concern over the availability of credit in rural areas led to the commissioning of the Vyas Committee that catalyzed a reform to rural branching and presumably the broader policy reform in 2005, early policy effects concentrated in agriculture could be expected. The analysis below indeed shows an early response in increased agricultural lending that wanes as the details of the reform become known, and increases again with greater emphasis placed on rural and semi-urban markets by later refinements to the reform. Attention is then turned to the effect of expanded banking services on agricultural performance.

1.6.2.1 Credit to Agriculture

Figure 1.14 shows the reduced form RD in the district percentage change in credit amount to direct and indirect agricultural activity in rural and semi-urban areas from their 2001 levels.[17] A positive and statistically significant response in under banked districts is first detected for credit to direct agricultural activities in 2005, the year following the Vyas Committee Report. The effect continues into 2006 after which it diminishes for a few years until emerging again in 2009. The greater relative expansion of credit in under banked districts at the cutoff is significant in magnitude as well, exceeding the sample means in 2005 and 2006, and remaining above 50% of the sample mean in the later years. The analogous effect estimated for lending to indirect agricultural activities is imprecisely estimated until 2009 after which the discontinuity is positive, statistically significant, and large in magnitude exceeding the sample mean. Results are reported in table 1.8.

The early expansion of credit beginning in 2005 is consistent with the timing of the Vyas Committee and emphasis placed on agricultural lending by policy makers, as well as with the incentives of private banks to increase lending to secure their market share of profitable loans in anticipation of increased competition. The decrease after 2006 may be attributable to banks learning that the branching policy reform was less directly tied to agricultural lending than initially anticipated. Alternatively, a subsidized credit program to farmers exclusively through public sector banks initiated around that time may have drawn away demand for private loans, washing out the private bank effect in direct agricultural credit.[18] The growth in agricultural lending after 2008 in under banked districts has several potential explanations. New branches opening as a result of the policy are growing in strength during these years. A refinement to the branching policy in 2008 made metropolitan areas ineligible for under banked status in 2008, creating greater incentive to move into lower populated

[17] The percentage change is approximated using the difference in logs of credit amounts from the 2001 reported levels.

[18] The Credit Subvention Scheme operated through NABARD and exclusively distributed through government sector banks was initiated in 2006-2007.

areas. The adoption of a new branching policy in 2010 that placed greater emphasis on rural and semi-urban branch entry in all districts with a bonus for under banked areas also explains the faster growth in the sample mean relative to the discontinuity for direct agricultural lending in 2010 and 2011. Alternatively, a reform to priority sector lending in 2007 also placed greater emphasis on agricultural lending. Required investment in the Rural Development Infrastructure Fund for failing to meet priority sector quotas, first coming due in the 2009 financial year, was accounted as indirect agricultural lending by banks. Finally, the categorization of loans by the RBI was revised in 2008, making direct comparisons by sector pre- and post- 2008 less accurate. Without finer data on loans, disentangling the exact causes is likely not possible.

Rainfall Annual rainfall is undoubtedly an important input for agricultural performance in India. Figure 1.15 presents estimated discontinuities in the district averaged percentage deviations of rainfall measures from their mean levels across the points of measure within a district. Since rainfall is random and unaffected by the policy reform at the cutoff or anywhere else, this analysis also serves as a falsification test of the RD design. As anticipated, rainfall does not show significant discontinuities at the cutoff. This suggests the response from credit and agricultural performance is not discontinuously effected by exogenous productivity shocks around the cutoff in the years considered.

Crops Figure 1.16 shows the reduced form regression discontinuity analysis for yield and output for two major crops in India, cotton and wheat. I present discontinuity analysis for crop yield (tonnes per hectare of cultivated land) and output (tonnes). Each specification controls for the district averaged percentage deviations of rainfall, district population and its square, and the 2001 pre-randomization value of the dependent variable. The analysis for the output is the most striking for cotton, while the effect on yield is greater for wheat. The response from these individual crop statistics suggests the branching policy reform positively impacted output and/or productivity. However, considering crops individually, and absent price data for the crop output, makes interpreting the results difficult. Not every district produces each crop, or is well suited for every type of agriculture. Farmers may be moving in or out of crops based on their prices. Yields may decrease if farmers enter high paying crop markets with plots of land poorly conditioned for those crops. Alternatively, yields may rise if farmers invest more in productive and profitable crops. Thus, a measure better incorporating the incentives faced by the farmer is needed.

To address these concerns, I compute an index of crop yields similar to that used in Jayachandran (2006). The index is constructed as a weighted average of crop yields for rice, wheat, jowar and groundnut, using crop revenue shares for the district as weights (see Data Appendix for details). I am able to construct the measure for the July-June years 2001-2002 to 2007-2008 from data on crop prices and production statistics collected at the district level. The price data for crops is available for a slightly smaller set of districts and generally restricted to crops for which the particular district produces in greater volumes. The index carries the added benefit, however, that a wider set of districts in India produce at least one of the crops in volume, meaning the set of districts through time will change less than considering output from a single crop. The results from the reduced form RD analysis are shown in the top panel of table 1.11. The estimates show positive discontinuities after 2005, though are estimated imprecisely except in 2008.

To estimate the effect of banking activity on average crop yield, I estimate a fuzzy RD of the crop yield index on total private sector credit accounts, instrumenting for credit accounts with the discontinuity. In the bottom panel, I present the fuzzy RD results for the pre-reform and post-reform periods. No effect is estimated in the pre-reform period. In the post reform period, I estimate an average effect of 0.03 with 10% significance, which may be interpreted as every thousand private bank accounts increases the crop yield by an average of 3%. This is a little less than the average effect of a positive rainfall shock, for rainfall above the 80th percentile for that district, on crop yield estimated in Jayachandran (2006).

1.6.3 Industrial Activities

Though the initial drive of the policy reform may have been to increase financial inclusion in low population areas and increase the credit flow to agriculture, many of the populated centers of under banked districts benefited from increased branch entry. This section investigates to what extent manufacturing enterprises benefited from the expanded bank presence by receiving loans and being able to invest in productive assets.

Credit to Manufacturing and Processing Figure 1.17 presents the reduced form RD effect for the percentage change in credit amount to manufacturing and processing. The effect after 2007 resembles the expansion of bank branches, with a steadily growing positive effect in under banked districts. Turning to the fuzzy RD results, the estimates are all quite imprecisely estimated. This may be due to the lumpiness with which capital is acquired in manufacturing, with large projects arriving infrequently per district but constituting a large share of credit. Multi-plant manufacturing enterprises may also have been able to secure funds for plants in unbranched areas through their headquarters. As such, large manufacturing investments may be less reliant on local branch access. While less response in credit appears to be earmarked for manufacturing, it is possible that credit lent as personal loans, that shows a strong response to the reform, may end up funding capital investments or freeing up other resources that then get invested in enterprises. To investigate this possibility, the next section examines input decisions from registered manufacturing firms.

1.6.3.1 Evidence from the ASI

In table 1.13 I present the results from difference in differences analysis using data from the ASI. The analysis uses years 1999-2010. In column (1) I estimate the effect on logged assets excluding land and inventory. The average treatment effect is positive but imprecisely estimated at a value of 17%. The effect on logged working capital, in column (2), is estimated at 0.264 with significance at the 10% level. The effect on the amount of outstanding loans held by the firm is estimated to increase 24% with statistical significance at the 10% level. Total investment increased by 19.7%, with statistical significance at the 10% level. The capital labor ratio is estimated to increase by 3.4 in response to the policy and is also estimated with precision at the 10% significance level. The sample mean for the under banked states sample was 10.88 post reform, making this a sizable effect. The estimates are quite robust to considering other ranges of years around the reform. In each regression I control for the rural status of the enterprise, the age of the plant as measured by years since opening, the number of total enterprises in the firm to which the enterprise belongs, the logged number of employees at the enterprise to control for size, and state fixed effects with state specific time trends. I exclude industry fixed effects as new NIC codes were adopted in 2008, potentially making some industry codings inconsistent through the time series. In practice, the inclusion of 3 digit NIC codes has little effect on the estimates.

The significant increase in loans carried by enterprises from under banked districts in the post reform years would indicate that the increased banking activity is finding its way to the industrial sector. The increases in working capital as well as total investments suggests firms are expanding the use of productive inputs with the expansion of credit. Further, the increase in the capital labor ratio is consistent with previously credit constrained firms making investments in capital as those constraints are relaxed with the inflow of new formal credit. These adjustments to the productive technologies of the firm are likely to result in changes in efficiency. If credit rationing resulted in the misallocation of credit, the expansion of credit may produce large impacts if it helps correct inefficient dispersions of marginal products of capital across firms.

1.6.4 Economic Growth and Light emitted at Night

The final analysis following the RD design examines discontinuities in changes of the emission of light into space at night. Henderson et al. (2012) established that so called "night-lights" provide a reliable proxy for economic growth under certain caveats. Important among these is the prescription to compare changes in light through time for one area to those in another, rather than comparing levels of light only across places or levels of light only across time. There are several reasons for this: the time series is composed of readings taken by different satellites in different blocks of years. The instruments between satellites vary, and their precision changes with age. The raw data are also processed in ways undisclosed to researches and vary across years. Part of this processing includes decisions on interpreting very low levels of light. Thus, any errors or idiosyncrasies generated by these processes get accentuated by differences in the degree of urbanization across locations.

This analysis accounts for these concerns by estimating the discontinuity in the difference of logged average district light since 2004. Thus, the dependent variable can be interpreted as the approximate percentage change in average light emitted in a location from its level in 2004. The RD compares these changes in estimating the discontinuity at the threshold. Figure 1.18 graphically reports the discontinuities estimated with a 2nd degree polynomial, which better captures the underlying data. Since the level of light is reported from measurements taken during the calendar year, 2005 is the first year with months under the enacted reform. Estimates are presented in Table (1.14). A slight negative discontinuity is estimated in the first year and is a small fraction of the average percentage change in light for districts in the sample. The discontinuity is small and positive again in 2006 though the average change in districts was negative overall. A positive jump in the discontinuity to 9.4% appears in 2007 and is estimated significantly at the 1% level, with the average change in light for districts in the sample increasing as well to 11.4%. A similar response is found in 2008 with 2009 showing low levels of light emitted in general for the sample around the threshold and a smaller discontinuity. The last three years show similar discontinuities in light to 2007 and 2008, with 2011 estimated with precision at the 10% confidence level.

To estimate the effect of expanding branch presence on the change in the amount of light emitted in districts at night, I perform a fuzzy RD of the change in light on private bank branches for the pre-reform period, which in this case is only 2005, and the post reform period constituting years 2006-2012 pooled together controlling for year fixed effects in addition to district population and its square. I run the estimation using local linear regressions because these better fit the bank branching data and offer a strong first stage. The pre-reform effect reported in the lower panel is negative and small, consistent with the reduced form estimate for 2005. The conventional estimate reported in the post reform column is estimated positive, but small and insignificant. However, the bias-corrected measure which accounts for a local quadratic estimate with a wider bandwidth, better capturing the quadratic relationship in the night light data, yields a positive and significant coefficient. This estimate is significant at the 1% level and has a value of 0.012. The coefficient may be interpreted as the average effect of each bank branch from the reform period is to increase nighttime light by 1.2%. Taking the estimated elasticity of nighttime light to GDP from Henderson et al. (2012) of 0.3, this implies that each bank branch raises local GDP by approximately 0.36%. The average increase in bank branches in the post reform period is estimated at approximately 5, implying the total effect was an average increase of local GDP in the districts by 1.79%.

1.6.5 Robust to NREGA

A competing explanation for the change in the spatial allocation of bank branches, increased banking activity, and subsequent responses in economic outcomes is the introduction of the Mahatma Gandhi National Rural Employment Guarantee Act (NREGA) that closely coincided in time with the branching policy reform. The act constitutes a public works program aimed at relieving poverty in rural areas by providing 100 days of guaranteed work to individuals from rural areas. The implementation of NREGA occurred in three stages, with 200 districts selected to begin the program in the fiscal year April 2006 through March 2007, with 130 new districts introduced in 2007-8 and the remaining 263 districts introduced in 2008-9. Zimmermann (2012) and Klonner and Oldiges (2014) analyze the effect of NREGA and provide background on the program. Of particular importance to the current analysis, NREGA benefits were distributed through bank accounts. One may conclude that this would increase the demand for formal banking, potentially increasing both the geographic reach and level of banking services. While likely true, to confound the current results there must also be a discontinuous break in the implementation of the program and disbursement of benefits at the "under banked" cutoff used for the regression discontinuity.

Districts were assigned to the various roll-out phases based on a composite index on district "backwardness" from the National Planning Commission (2003). In table 1.15 I test whether a discontinuity in phase assignment can be detected at the cutoff. A significant discontinuity would suggest a correlation with the NREGA program. The test fails to reject the null hypothesis of continuity at the cutoff for all three phases. Thus, NREGA phase assignment and therefore likely its benefits as well, would be unexpected to differ at the cutoff. In analysis not shown, I perform a visual RD of the district composite index at the under banked cutoff. No discontinuity is observable at the cutoff. Further, the general notion that persons per branch is generally increasing with worsening district conditions is confirmed by the trend of the index on "backwardness."[19]

1.7 Conclusions

Greater access to formal financial markets through policy driven branch expansion can relax credit constraints, allowing productive firms to invest at their optimal rates and households to smooth consumption across time. However, due to information asymmetries, possibly misaligned incentives within bank hierarchies and high costs to serving certain areas, expanding branch networks need not necessarily result in greater credit access or economic growth. Identifying the effect of formal bank access also proves challenging due to classic endogeneity concerns of branch location choice. This paper utilizes exogenous variation

[19] Out of concern that the omitted districts are disproportionately from one side of the cutoff or the other, I repeat the McCrary test only including districts missing the composite index value. I fail to reject the null hypothesis of continuity in the density of districts at the cutoff with the discontinuity estimate in the log difference in height at -31 and a standard error of 38.

in formal banking access generated by a recent and previously unstudied branching policy reform in India from 2005. Using a regression discontinuity design based on the assignment rule of districts to under banked status, the analysis demonstrates an expansion of banking services consistent with the timing and incentives of the reform. Agricultural productivity and the capital intensity of manufacturing are shown to increase in areas receiving higher credit due to the reform. I estimate that an increase of 1,000 private bank credit accounts in a district raises average crop yield by 3%. This effect is a little less than half of the effect Jayachandran (2006) measures on crop yield from positive rainfall shocks. Manufacturing enterprises in areas with increased access to banking exhibited higher growth in total investments, working capital and capital labor ratios. Finally, I confirm the aggregate effect on local GDP growth by showing that areas with expanding banking services experienced higher rates of growth in nighttime light intensity in the years following the reform. The estimates imply that each additional private bank branch led to a 0.36% increase in local GDP.

Overall, these findings offer strong causal evidence that the expansion of the financial systems facilitate growth in productive activities important for driving economic development. They further confirm the potential effectiveness of policy reforms in producing this expansion. The evidence suggests that the competition anticipated to be generated by the reform led profit maximizing banks to increase the quantity of credit ahead of additional branch entry to secure market shares in profitable areas. This effect from competition enhanced the response to the reform, beyond what could be expected from branch entry alone. Importantly, the expansion of credit does not seem to have been restricted to urban areas which is a common concern in developing areas. Rural and semi-urban markets in underserved areas also exhibited increases of credit from private banks. Though caution must be used in examining the credit data at disaggregated levels, the timing and effects in rural areas are consistent with banks responding to priority sector requirements by lending to target groups through their newly opened branches. This would suggest the importance of harmonizing incentives across policies to effectively meet policy objectives. Future work should examine the effects of interactions between the policy interventions.

The results of this analysis suggests that private sector banks respond strongly to the incentives generated by regulations. Areas targeted by the policy reform demonstrated a variety of benefits. While this speaks to the effects of financial development, little can be said in terms of overall policy evaluation. Resources were redistributed across districts due to the reform. Without measures for the opportunity costs of these resources, a full welfare analysis is beyond the scope of this paper. The introduction of bank or branch level data on credit, as well as farmer level data of characteristics, borrowing, inputs and crop selection would allow for a closer examination of these potential costs, give greater insights as to the specific mechanisms driving productivity gains and provide a basis for a more complete welfare analysis. Future work should also concentrate on a deeper understanding of the interactions between private sector, public sector and informal lenders in facilitating financial development.

1.8 Figures

Figure 1.1: Banking Sector Structure in India

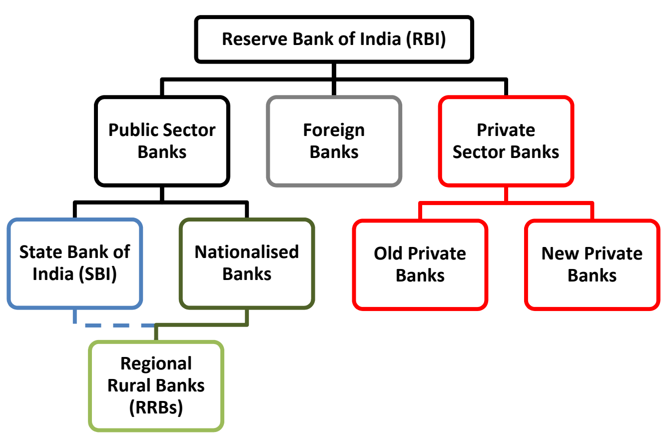


Figure 1.2: Policy Time Line for Bank Branching and Related Reforms

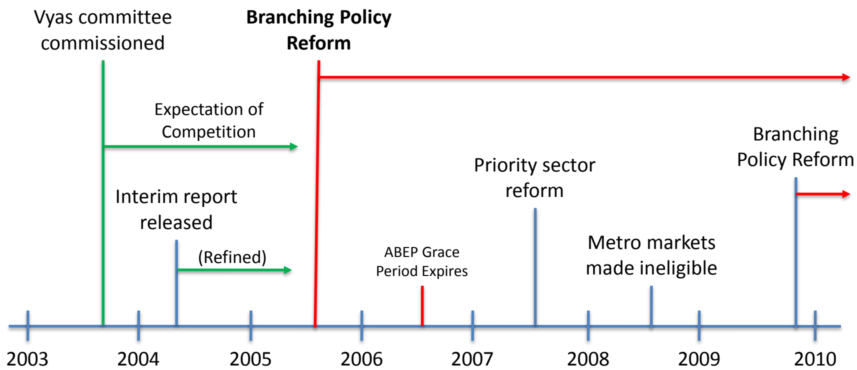
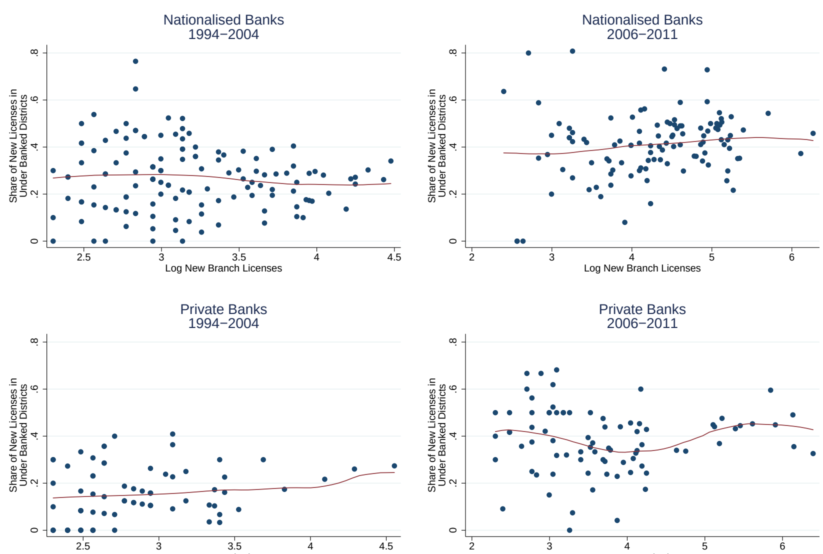
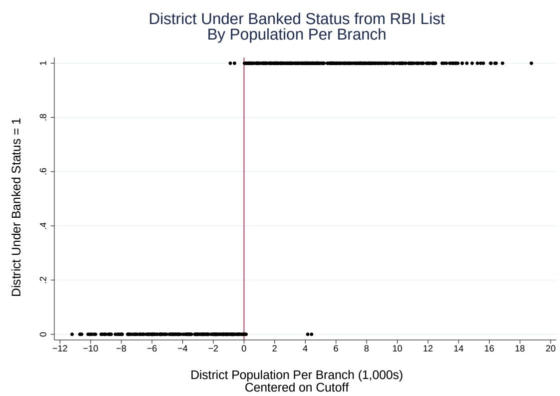


Figure 1.3: Bank-year Specific Shares of New Licenses Issued to the Bank in Districts on the 2006 List of Under Banked Districts



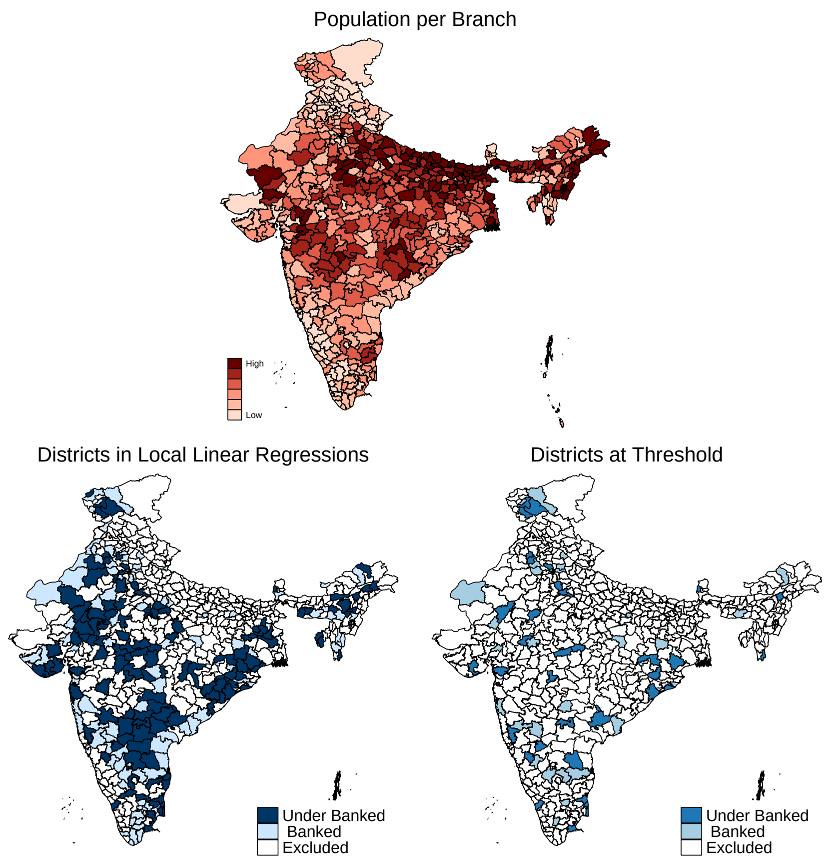
Note: Each observation is the percentage of new licenses for general and specialized branches issued during a calendar year for each bank-year pair in the respective Nationalised and Private bank groups. The value is plotted against the log of total new licenses issued to the bank in that year.

Figure 1.4: RD Visual First Stage: Under Banked Status by District Population Per Branch



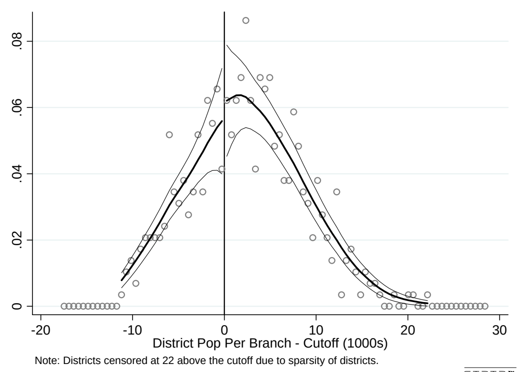
Note: The dots report the under banked status of a district, taking a value equal to one if the district appeared on the list of under banked districts in the 2006 RBI MC on Branching Authorisation Policy, and zero otherwise. The forcing variable, district population per branch centered on the national average, is on the x-axis scaled to thousands of persons per branch. Values to the right of the cutoff are predicted to have under banked status. 368 districts of 572 have under banked status, with 6 incorrect predictions based on the rule.

Figure 1.5: Maps of Under Served Areas by Formal Banking



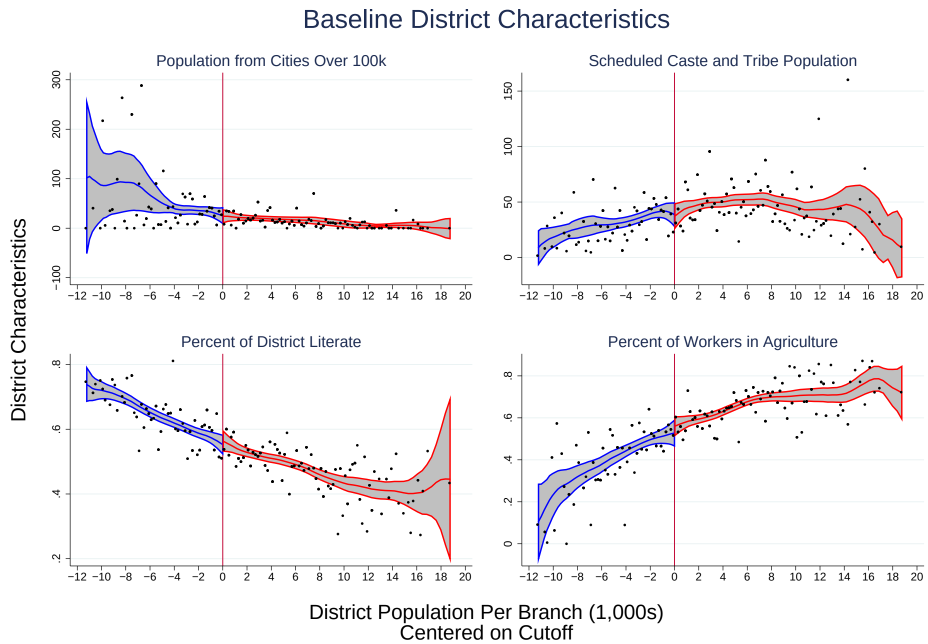
Upper: District population per branch. Lower Left: Including only districts used in local linear regressions. Lower Right: Including only those districts within 1,000 persons per branch of the cutoff.

Figure 1.6: Visual McCrary Test



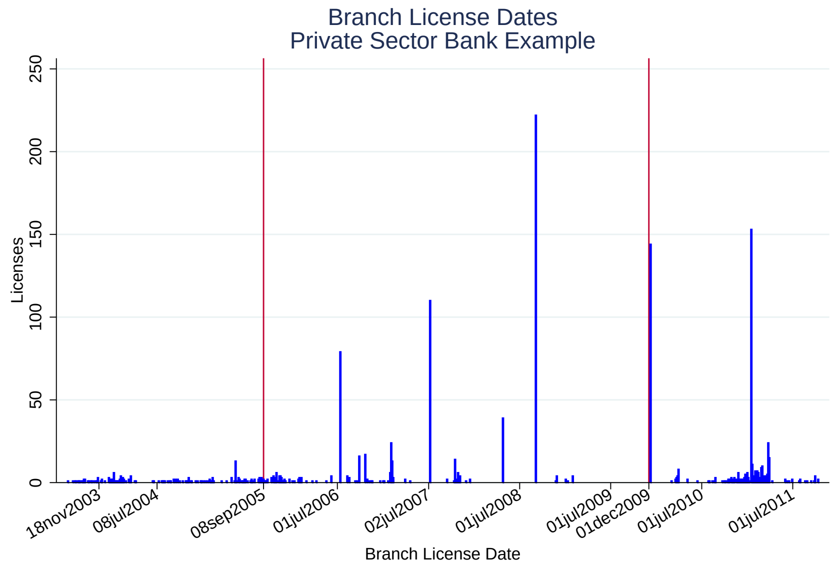
The graph plots a density of districts along the forcing variable, district population per branch, centered on the cutoff. The discontinuity estimate in the log difference in height is 6.6 with a standard error of 22. I fail to reject the null hypothesis of continuity at the cutoff, suggesting a lack of manipulation.

Figure 1.7: Continuity Around the Threshold



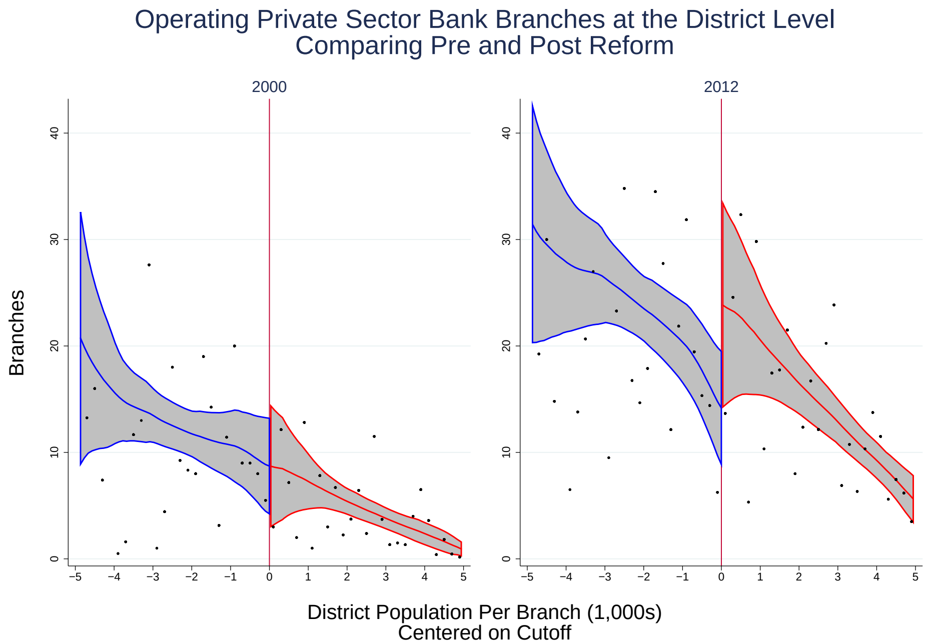
Note: The figure presents baseline district characteristics taken from the 2001 Population Census of India, with dots reporting local averages for districts falling within non-overlapping 200 persons per branch bins. The horizontal axis is the forcing variable of district population per branch centered on the cutoff. Districts predicted to have under banked status fall to the right of the cutoff. The estimated y-value from a local linear regression of bandwidth 3.5 thousand persons per branch is shown at each x-value, allowing for different slopes on either side of the cutoff, with 5% confidence intervals.

Figure 1.8: Histogram of Branch Licenses Showing ABEPs for a Large Private Sector Bank



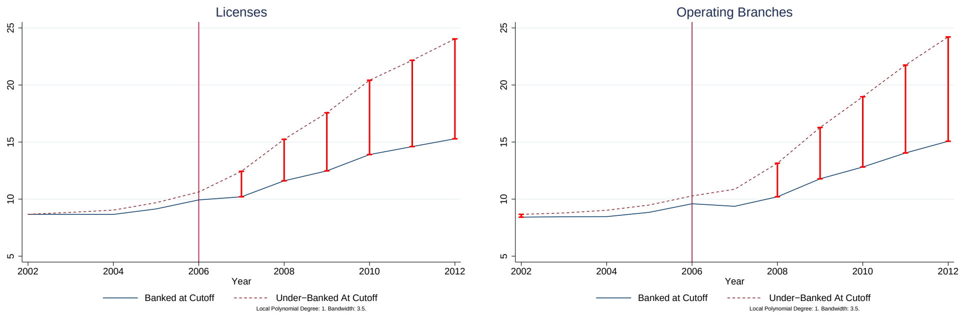
Note: Branch license dates are from the MOF. Bin widths are set to 4 days. Though annual branch expansion plans (ABEP) may not be observed directly, the large spikes in branch licenses set approximately a year apart after 2005 are consistent with licenses issued through ABEPs. The dates of Master Circular releases are shown, with vertical red lines at the 2005 policy reform and the subsequent reform in December 2009. Branches acquired through mergers and acquisitions are excluded.

Figure 1.9: Visual RD: Operating Private Bank Branches



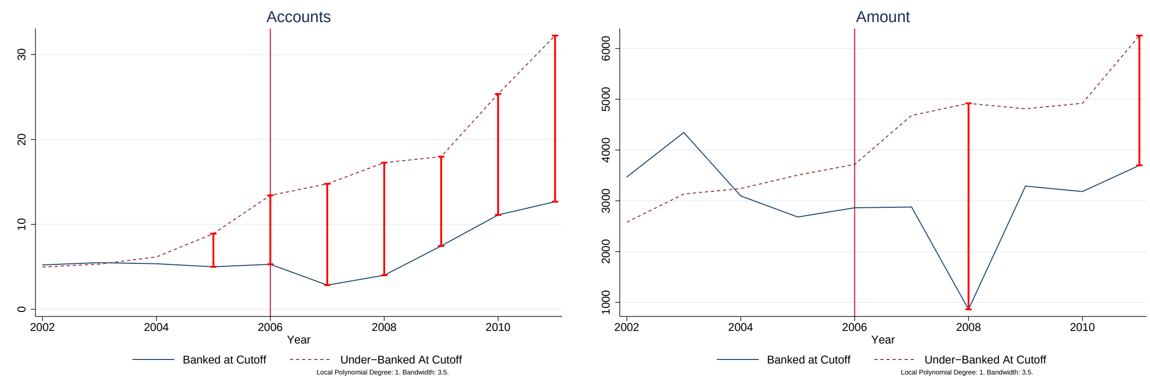
Note: Each plot presents the number of operating private sector bank branches within a district, in respective years, with dots reporting local averages for districts falling within non-overlapping 200 persons per branch bins. The horizontal axis is the forcing variable of district population per branch centered on cutoff and scaled to thousands of persons per district. The estimated y-value from local linear regressions, with a 3.5 thousand persons per district bandwidth and triangular kernel, at each x-value along with 5% confidence intervals is shown, allowing for different slopes on either side of the cutoff. The year 2000 in the left plot shows a pre-reform example of branches around the cutoff. The figure on the right shows the cumulative effect of the policy on operating branches since its implementation in 2005. Local averages greater than 40 are not shown in the plots, but were included in local linear regressions. Local averages greater than 40 did not occur close to the cutoff.

Figure 1.10: Discontinuity from Reduced Form: Operating Private Bank Branches



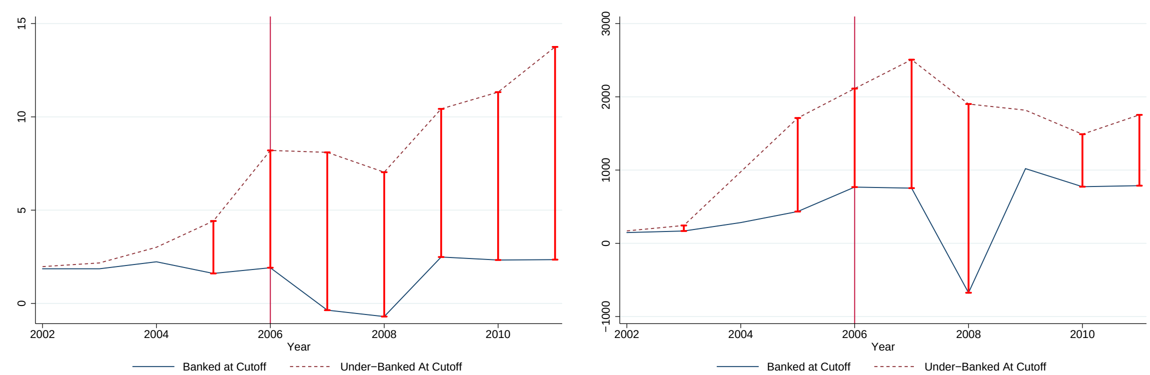
Note: Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel. The figure plots the estimated intercepts at the cutoff from the estimation of the RD equation repeated annually. The red dashed line provides the estimated intercept from approaching the threshold along the under banked side. The solid blue line reports the corresponding intercept approaching from the banked side. The distance between the two, reported for each year, shows the estimated discontinuity at the threshold. A vertical red line between the two points indicates a positive discontinuity with under banked districts exhibiting a higher value at the threshold than banked districts, with significance at least at the 10% level. A vertical dashed green line indicates a negative discontinuity estimated at least at the 10% level. The thin vertical red line at 2006 represents the first estimation made after the reform implementation.

Figure 1.11: Discontinuity from Reduced Form: Private Banks Aggregate Credit



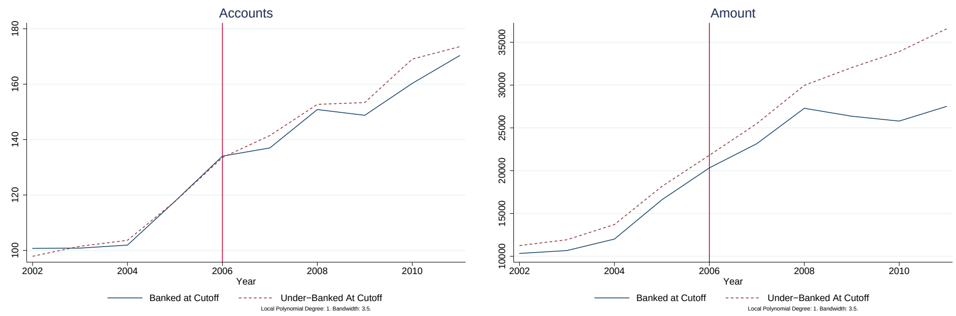
Note: Accounts reported in thousands. Amounts reported in millions of rupees. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel. See notes from Table 1.10 for graph description.

Figure 1.12: Discontinuity from Reduced Form: Private Credit to Personal Loans



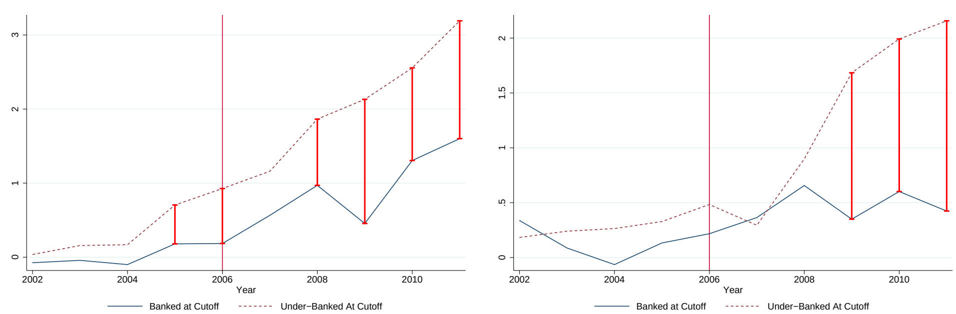
Note: Accounts reported in thousands. Amounts reported in millions of rupees. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel. See notes from Table 1.10 for graph description.

Figure 1.13: Discontinuity from Reduced Form: Credit from Public Sector Banks



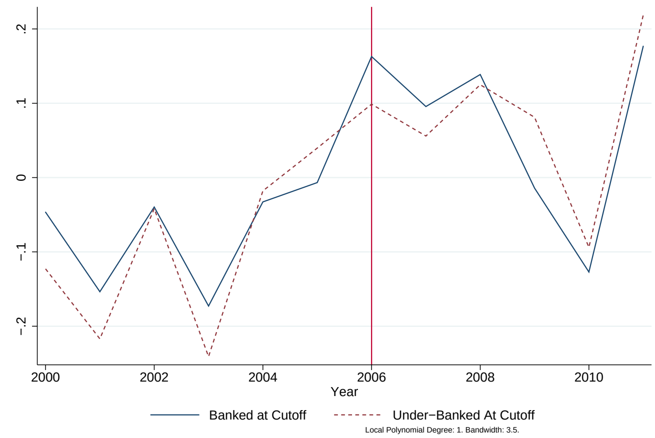
Note: Accounts reported in thousands. Amounts reported in millions of rupees. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel. See notes from Table 1.10 for graph description. Public sector banks include State Bank of India and Associated Banks, Nationalised Banks, IDBI and Regional Rural Banks.

Figure 1.14: Discontinuity from Reduced Form: Percentage Change in Private Credit Amount to Agriculture in Rural and Semi-Urban Areas



(Left) Direct Agriculture, (Right) Indirect Agriculture. Note: Percentage change is approximated using difference in logs relative the value reported in 2001. Estimated using local linear regressions with controls for district population and its square. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel.

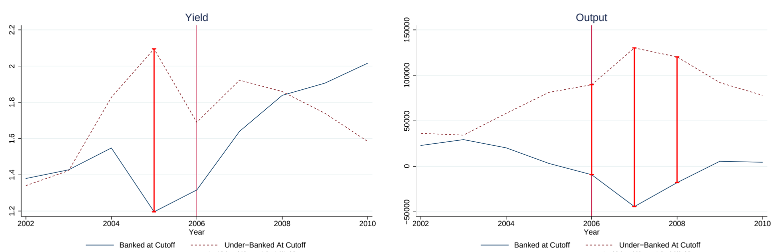
Figure 1.15: Discontinuity from Reduced Form: Rainfall



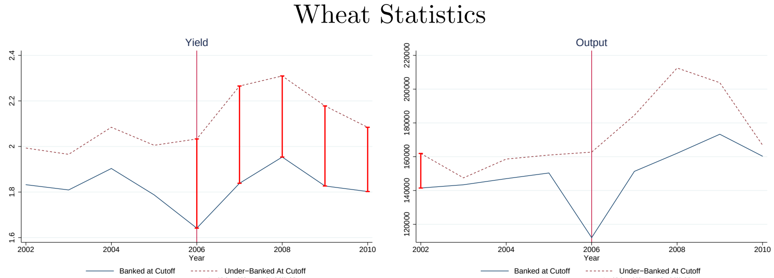
Note: District Average Percentage Deviation from Mean. Estimated using local linear regressions. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel. See notes from Table 1.10 for graph description.

Figure 1.16: Discontinuity from Reduced Form: Individual Crops

Cotton Statistics

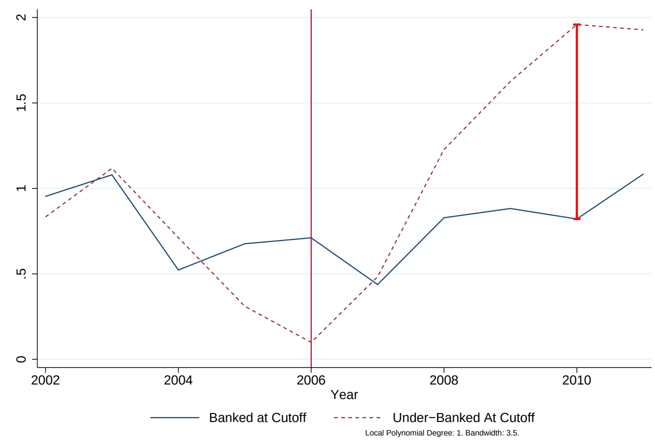


Wheat Statistics



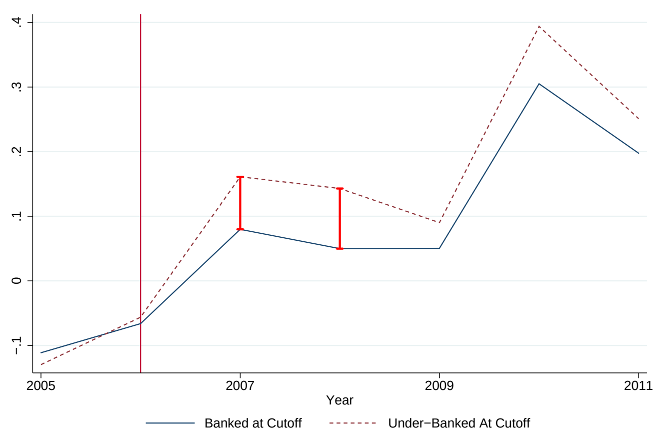
Note: Yield [Tonnes/Hectare] (Left), Output [Tonnes] (Right). Cotton output measured in bales rather than tonnes.

Figure 1.17: Discontinuity From Reduced Form: Percentage Change in Private Credit Amount to Manufacturing and Processing from 2001 Level



Note: Percentage change is approximated using difference in logs relative the value reported in 2001. Estimated using local linear regressions with controls for district population and its square. Bandwidths are set 3.5 thousand persons per branch and estimated using a uniform kernel.

Figure 1.18: Discontinuity from Reduced Form: Difference in Log Mean District Light from 2004 Level



Note: Estimated using local quadratic regressions with controls for district population and its square. Bandwidths are set to 3.5 thousand persons per branch and estimated using a triangular uniform.

1.9 Tables

Table 1.1: Continuity tests for Baseline Values at the Cutoff

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Population	City_Pop	Sched_Caste_Tribe_Pop	Pct_Literate	Pct_Dist_Dark	Area_Proxy	PrivBranches2000
Conventional	0.839	-1.344	-1.436	0.0114	-0.00894	-2,485	0.192
	[35.38]	[13.61]	[8.483]	[0.0219]	[0.0169]	[2,697]	[3.026]
Bias-corrected	16.01	2.353	0.265	0.0187	-0.0101	-3,386	0.567
	[35.38]	[13.61]	[8.483]	[0.0219]	[0.0169]	[2,697]	[3.026]
Robust	16.01	2.353	0.265	0.0187	-0.0101	-3,386	0.567
	[42.75]	[16.30]	[9.840]	[0.0261]	[0.0210]	[3,323]	[3.527]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	95	95	95	95	95	95
N_UBanked	122	122	122	122	122	122	122
DepMean	176.7	28.56	45.24	0.553	0.949	8150	7.198

Standard errors in brackets

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Note: Estimated using local linear regressions with no controls. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel.

Table 1.2: Summary Statistics

Banking

	Banked, Pre-reform			Banked, Post-reform			Under Banked, Pre-reform			Under Banked, Post-Reform		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Branches												
SBI	610	28.618	23.095	732	33.238	27.971	900	21.35	16.381	1080	24.595	19.693
Nationalised	610	69.805	62.759	732	80.634	73.984	900	45.444	44.86	1080	51.54	50.432
RRB	610	21.523	21.684	732	23.001	22.171	900	28.221	22.147	1080	29.207	22.946
Foreign	610	0.121	0.624	732	0.243	1	900	0.018	0.199	1080	0.112	0.457
Old Private	610	11.807	16.582	732	11.628	15.295	900	4.198	9.298	1080	4.589	10.001
New Private	610	2.428	4.6	732	7.25	10.687	900	0.794	2.372	1080	4.049	6.154
Public Banks	610	120.154	87.587	732	137.926	105.531	900	95.064	66.491	1080	106.006	76.429
Private Banks	610	14.234	18.58	732	18.878	20.755	900	4.992	10.375	1080	8.638	13.96
Credit Amount												
SBI	610	5293.635	5980.068	732	11037.746	12248.838	900	3285.651	5986.45	1080	6507.465	8548.992
Nationalised	610	10236.988	13154.392	732	22228.233	33180.444	900	4602.575	5692.052	1080	9362.257	12494.337
RRB	610	870.748	1198.64	732	1738.277	2270.793	900	950.135	1134.78	1080	1869.281	2256.909
Foreign	610	201.344	727.787	732	487.36	1620.559	900	50.173	293.389	1080	191.788	1414.19
Private	610	3813.913	7071.325	732	7637.427	12055.826	900	1354.922	3542.466	1080	2437.963	5464.27
Credit Accounts												
SBI	610	30945.372	31517.419	732	47639.104	50181.875	900	24107.006	24218.304	1080	38046.444	39105.763
Nationalised	610	60582.561	60584.955	732	89278.02	97041.327	900	37963.999	38526.215	1080	55938.739	58976.202
RRB	610	22255.538	33920.327	732	30088.209	47295.116	900	28251.067	34646.607	1080	36354.233	48093.88
Foreign	610	134.425	772.631	732	319.858	1656.413	900	51.02	564.603	1080	119.098	874.722
Private	610	9792.657	14751.414	732	25507.242	35027.737	900	3214.418	7356.894	1080	9889.303	22363.595
Deposit Amount												
SBI	607	9599.797	10660.293	732	16412.707	20661.421	892	6104.533	6197.594	1078	10180.87	10886.087
Nationalised	607	20027.738	26126.927	732	33469.464	51159.493	892	9745.183	12975.665	1078	15306.32	20677.413
RRB	607	1340.932	1519.9	732	2212.508	2520.006	892	1807.669	1792.853	1078	2828.679	2818.4
Foreign	607	181.203	1207.168	732	611.752	4849.064	892	20.185	243.413	1078	65.089	603.547
Private	607	4695.24	8722.103	732	8973.14	17799.643	892	1371.376	2938.947	1078	2798.099	5257.67
Deposit Accounts												
SBI	607	203.438	178.676	732	298.246	276.023	892	147.726	130.511	1078	232.131	228.963
Nationalised	607	502.83	502.301	732	683.751	696.657	892	294.637	342.959	1078	410.485	464.146
RRB	607	76.55	101.221	732	118.76	157.796	892	100.515	109.819	1078	157.422	174.789
Foreign	607	0.98	6.606	732	2.268	14.987	892	0.188	2.342	1078	0.396	2.834
Private	607	91.003	124.465	732	136.977	184.145	892	30.155	63.394	1078	50.568	93.778

Source: RBI Master Office File, BSR 1 and BSR 2 years 2001-2011. Sample includes years 2001-2011 for districts falling within 5 thousand persons per branch of the national average. Each year includes 122 banked districts and 180 under banked districts, from a total of 572 districts considered. Amounts are reported in Rupees million adjusted to 2011q4 prices; Accounts are reported in thousands.

Table 1.3: Summary Statistics Continued...

Agriculture

	Banked, Pre-reform			Banked, Post-reform			Under Banked, Pre-reform			Under Banked, Post-reform		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Cotton												
Area	403	32,656	53,321	349	31,406	56,677	619	31,351	64,876	471	37,076	75,472
Output	403	59,959	127,462	349	100,347	229,598	619	41,581	89,199	471	86,119	203,562
Productivity	403	1.61	0.98	349	2.12	1.38	619	1.35	0.84	471	1.55	1.28
Maize												
Area	560	11,945	20,923	470	15,124	28,518	968	16,400	32,962	761	16,688	36,426
Output	560	27,988	57,175	470	48,069	103,819	968	28,449	64,162	761	34,070	87,053
Productivity	560	1.87	1.19	470	2.38	2.24	968	1.49	0.84	761	1.76	1.35
Onion												
Area	431	1,527	3,714	342	2,036	5,455	743	1,074	2,489	510	1,485	4,019
Output	431	13,885	29,608	342	17,539	36,355	743	14,587	51,185	510	24,189	99,249
Productivity	431	11.71	7.93	342	12.03	8.58	743	11.34	7.48	510	11.38	7.92
Potato												
Area	351	2,028	4,026	303	2,303	6,024	674	3,014	9,512	587	3,694	12,041
Output	351	28,503	44,128	303	27,843	43,051	674	67,058	248,196	587	71,627	286,377
Productivity	351	13.75	7.51	303	12.93	7.79	674	12.64	7.55	587	11.76	8.19
Rice												
Area	667	64,626	82,739	544	67,299	85,705	1017	88,839	104,258	784	100,968	120,405
Output	667	173,077	285,059	544	194,407	303,283	1017	160,160	221,919	784	197,829	266,243
Productivity	667	2.30	1.01	544	2.51	1.10	1017	1.61	0.87	784	1.81	0.94
Sesamum												
Area	573	3,245	6,935	460	2,790	4,742	908	4,826	11,359	749	5,919	15,535
Output	573	1,220	3,198	460	1,119	2,212	908	1,805	5,529	749	2,032	6,103
Productivity	573	0.35	0.23	460	0.38	0.25	908	0.32	0.22	749	0.35	0.24
Sugarcane												
Area	523	12,161	23,096	419	11,554	22,413	907	8,554	25,972	711	8,866	27,790
Output	523	955,008	1,797,426	419	902,855	1,738,094	907	590,206	1,786,733	711	588,924	1,878,506
Productivity	523	70.26	35.51	419	67.35	39.47	907	53.13	26.72	711	55.86	30.25
Tobacco												
Area	166	7,958	16,242	176	8,267	17,829	258	454	1,647	213	620	2,082
Output	166	9,853	22,353	176	10,113	20,766	258	663	2,233	213	1,128	3,622
Productivity	166	1.54	1.53	176	1.53	1.61	258	1.63	1.88	213	1.71	1.57
Wheat												
Area	437	60,088	81,807	349	64,550	81,240	923	49,803	65,451	689	52,869	67,471
Output	437	204,344	353,065	349	225,183	353,261	923	126,363	200,516	689	147,671	224,604
Productivity	437	2.21	1.25	349	2.38	1.27	923	1.78	0.97	689	1.93	1.02

Source: Rainfall data from TRMM satellite, crop data from State Agricultural Reports. Sample includes years 2000-2010 for districts falling within 5 thousand persons per branch of the national average. Observations are crop-years; the number of districts varies by crop as not every crop is grown in all districts. 302 of 572 districts are eligible for sample. Area is reported in Hectares square, output in tonnes, and productivity is output divided by area. Cotton reported in bales instead of tonnes.

Annual Survey of Industries

	Banked, Pre-reform			Banked, Post-reform			Under Banked, Pre-reform			Under Banked, Post-reform		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Log Total Employees	42702	3.786	1.403	40252	3.954	1.436	21133	3.567	1.345	17976	3.72	1.403
Log Number of units	42824	0.04	0.193	40575	0.041	0.203	21216	0.021	0.15	18123	0.025	0.159
Plant Age	42248	16.002	13.986	39268	15.204	13.878	20864	14.97	14.197	17562	14.664	14.332
Log Capital												
(No Land or Inventory)	42339	14.911	2.876	39707	15.151	3.392	21030	14.576	2.952	17886	14.995	3.135
Log Net Assets	42352	15.679	2.883	39772	15.76	3.294	21040	15.354	2.929	17902	15.602	3.024
Log Working Capital	35823	15.306	3.024	34057	15.259	3.689	18262	15.015	3.105	15818	15.287	3.154
Log Loans	34828	14.869	4.037	32543	14.962	4.199	16258	14.874	4.084	13795	15.062	4.035
Log Total Investment	39950	14.688	3.2	37858	14.943	3.829	20517	14.248	3.298	17468	14.649	3.619
Capital Labor Ratio	42221	6.644	47.52	39543	11.121	237.379	20971	8.133	38.898	17800	10.879	105.471
Log Capital Labor Ratio	42202	0.774	1.535	39535	0.875	1.516	20958	0.89	1.662	17798	1.003	1.645

Source: Annual Survey of Industries, Unit level data 1999-2010. Sample is restricted to plants reporting being open and reporting a valid urban or rural status. Captital Labor Ratio constructed as average of opening and closing Net Assets divided by the total wage bill plus benefits. States and UTs selected by their share of population being concentrated on one side of the threshold or the other. "Banked States" include Haryana, Uttarakhand, Punjab, Mizoram, Daman and Dimiu, Karnataka, Puducherry, and "Under Banked States" include Rajasthan, Tripura, Jharkhand, Orissa, Dadra and Nagar Haveli.

Table 1.4: Fuzzy RD: Private Bank Branches

Licenses

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011	(11) 2012
Conventional	-0.0469	0.246	0.414	0.533	0.657	2.198*	3.785**	5.691**	7.313**	8.510***	9.665***
	[0.182]	[0.323]	[0.533]	[0.693]	[0.911]	[1.212]	[1.668]	[2.267]	[2.869]	[3.267]	[3.545]
Bias-corrected	-0.0888	0.318	0.869	0.936	1.067	2.628**	4.153**	6.265***	7.748***	8.910***	10.12***
	[0.182]	[0.323]	[0.533]	[0.693]	[0.911]	[1.212]	[1.668]	[2.267]	[2.869]	[3.267]	[3.545]
Robust	-0.0888	0.318	0.869	0.936	1.067	2.628*	4.153**	6.265**	7.748**	8.910**	10.12**
	[0.215]	[0.379]	[0.643]	[0.831]	[1.075]	[1.423]	[1.939]	[2.638]	[3.342]	[3.796]	[4.123]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	122	122	122	122	122	122	122	122	122
DepMean	8.714	8.917	9.241	9.847	10.62	11.92	13.83	15.31	17.13	18.47	19.99

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Operating Branches

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011	(11) 2012
Conventional	0.155	0.320	0.538	0.579	0.629	1.128	3.017**	4.512**	6.630***	8.680***	9.991***
	[0.159]	[0.329]	[0.561]	[0.669]	[0.902]	[1.070]	[1.445]	[1.962]	[2.541]	[3.184]	[3.610]
Bias-corrected	0.133	0.355	0.990*	0.958	1.032	1.298	3.463**	4.770**	6.919***	9.097***	10.41***
	[0.159]	[0.329]	[0.561]	[0.669]	[0.902]	[1.070]	[1.445]	[1.962]	[2.541]	[3.184]	[3.610]
Robust	0.133	0.355	0.990	0.958	1.032	1.298	3.463**	4.770**	6.919**	9.097**	10.41**
	[0.186]	[0.383]	[0.668]	[0.801]	[1.069]	[1.250]	[1.674]	[2.288]	[2.961]	[3.705]	[4.197]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	122	122	122	122	122	122	122	122	122
DepMean	8.636	8.801	9.125	9.597	10.34	10.87	12.25	14.42	16.19	17.91	20.00

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel. Under banked status is instrumented for with predicted under banked assignment. Licenses are considered in operation if they are granted for a branch currently operating or pending opening.

Table 1.5: Fuzzy RD: Private Banks Aggregate Credit

Private Sector Credit Accounts

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011
Conventional	-0.264	0.119	1.147	3.678*	6.725*	10.42***	10.01*	7.781	11.74*	16.96
	[0.606]	[0.857]	[1.270]	[1.993]	[4.039]	[3.698]	[5.461]	[4.872]	[7.009]	[10.95]
Bias-corrected	-0.281	0.523	1.871	4.199**	8.192**	12.14***	11.90**	9.321*	13.44*	18.26*
	[0.606]	[0.857]	[1.270]	[1.993]	[4.039]	[3.698]	[5.461]	[4.872]	[7.009]	[10.95]
Robust	-0.281	0.523	1.871	4.199*	8.192*	12.14***	11.90*	9.321	13.44	18.26
	[0.707]	[1.023]	[1.489]	[2.371]	[4.741]	[4.549]	[6.439]	[5.692]	[8.198]	[12.73]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	5.067	5.484	6.470	8.800	12.83	13.77	16.78	17.78	22.82	25.80

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Private Sector Credit Amounts

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011
Conventional	-550.7	-867.3	-159.9	994.4	1,060	1,664	3,504	1,451	1,461	2,194
	[1,547]	[2,026]	[1,947]	[1,342]	[1,682]	[2,156]	[2,529]	[1,475]	[1,363]	[1,646]
Bias-corrected	-182.5	-359.5	294.2	1,813	1,893	2,523	4,854*	2,703*	2,314*	3,044*
	[1,547]	[2,026]	[1,947]	[1,342]	[1,682]	[2,156]	[2,529]	[1,475]	[1,363]	[1,646]
Robust	-182.5	-359.5	294.2	1,813	1,893	2,523	4,854	2,703	2,314	3,044
	[1,760]	[2,326]	[2,221]	[1,598]	[2,072]	[2,623]	[3,314]	[1,879]	[1,679]	[2,054]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	2641	3223	2943	3466	3922	4920	5934	5362	4932	5990

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Accounts reported in thousands. Amounts reported in millions of rupees. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel. Under banked status is instrumented for with predicted under banked assignment.

Table 1.6: Fuzzy RD: Private Credit to Personal Loans

Accounts in Personal Loans

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Conventional	0.147	0.460	1.109	2.898**	6.498**	8.159***	6.706**	6.957**	8.658***	10.89*
	[0.240]	[0.447]	[0.816]	[1.216]	[2.647]	[2.455]	[2.732]	[3.124]	[3.047]	[5.652]
Bias-corrected	0.205	0.658	1.473*	3.127**	7.307***	9.538***	7.775***	7.839**	9.870***	11.98**
	[0.240]	[0.447]	[0.816]	[1.216]	[2.647]	[2.455]	[2.732]	[3.124]	[3.047]	[5.652]
Robust	0.205	0.658	1.473	3.127**	7.307**	9.538***	7.775**	7.839**	9.870***	11.98*
	[0.285]	[0.518]	[0.945]	[1.486]	[3.140]	[3.179]	[3.359]	[3.707]	[3.614]	[6.562]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	1.885	1.910	2.686	3.681	6.138	6.313	6.471	9.120	9.056	9.707

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Amount to Personal Loans

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Conventional	22.11	75.08	663.6	1,382**	1,605**	1,927**	2,409	628.1	692.1	878.2*
	[20.50]	[51.87]	[480.0]	[657.6]	[753.1]	[887.3]	[1,533]	[626.7]	[425.4]	[485.3]
Bias-corrected	33.25	103.3**	759.6	1,499**	1,846**	2,179**	3,008**	687.7	807.5*	925.9*
	[20.50]	[51.87]	[480.0]	[657.6]	[753.1]	[887.3]	[1,533]	[626.7]	[425.4]	[485.3]
Robust	33.25	103.3*	759.6	1,499**	1,846**	2,179**	3,008	687.7	807.5	925.9*
	[23.46]	[60.87]	[497.4]	[763.0]	[901.2]	[1,065]	[2,151]	[735.6]	[493.8]	[556.7]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	151.8	199.6	644.3	1003	1384	1658	1983	1609	1200	1280

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Accounts reported in thousands. Amounts reported in millions of rupees. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel. Under banked status is instrumented for with predicted under banked assignment.

Table 1.7: RD from Reduced Form: Credit from Public Sector Banks

Public Sector Credit Accounts

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011
Conventional	-3.827	-0.621	0.269	-1.419	-1.086	4.170	-0.156	2.630	4.258	-1.188
	[3.191]	[4.109]	[5.598]	[7.397]	[10.56]	[10.96]	[11.23]	[12.24]	[15.05]	[15.09]
Bias-Corrected	-3.866	-0.196	1.956	1.412	1.662	8.924	2.854	7.809	8.348	2.285
	[3.191]	[4.109]	[5.598]	[7.397]	[10.56]	[10.96]	[11.23]	[12.24]	[15.05]	[15.09]
Robust	-3.866	-0.196	1.956	1.412	1.662	8.924	2.854	7.809	8.348	2.285
	[3.814]	[5.045]	[6.903]	[9.092]	[12.92]	[13.43]	[13.69]	[14.88]	[17.98]	[18.26]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	94	94	94	94	94	94	93	94	94	94
N_UBanked	122	122	122	122	122	122	122	122	122	122
DepMean	99.43	102.5	105.6	120.6	132.4	141.5	151.7	154.2	167.2	177.1

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Public Sector Credit Amounts

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011
Conventional	585.4	791.0	795.6	252.2	-669.7	27.51	-849.1	2,754	2,948	3,923
	[528.0]	[774.0]	[982.6]	[2,482]	[3,326]	[3,029]	[4,649]	[3,457]	[3,487]	[3,949]
Bias-Corrected	531.9	1,086	1,073	639.5	-421.5	865.7	322.0	4,478	4,171	5,329
	[528.0]	[774.0]	[982.6]	[2,482]	[3,326]	[3,029]	[4,649]	[3,457]	[3,487]	[3,949]
Robust	531.9	1,086	1,073	639.5	-421.5	865.7	322.0	4,478	4,171	5,329
	[599.5]	[1,058]	[1,343]	[2,974]	[3,959]	[3,546]	[5,400]	[4,124]	[4,329]	[4,995]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	94	94	94	94	94	94	93	94	94	94
N_UBanked	122	122	122	122	122	122	122	122	122	122
DepMean	10544	11953	13493	17693	21386	23326	27547	29581	31372	34125

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note: Accounts reported in thousands. Amounts reported in millions of rupees. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel. Public sector banks include State Bank of India and Associated Banks, Nationalised Banks, IDBI and Regional Rural Banks.

Table 1.8: Fuzzy RD: Percentage Change in Private Credit Amount to Rural and Semi-Urban Areas

Direct to Agriculture

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011
Conventional	0.0681	0.141	0.216	0.699**	1.033**	0.734	0.945	1.795***	1.407**	1.731***
	[0.108]	[0.166]	[0.209]	[0.343]	[0.418]	[0.534]	[0.666]	[0.679]	[0.686]	[0.659]
Bias-corrected	0.120	0.266	0.301	0.917***	1.284***	0.899*	1.236*	2.104***	1.719**	1.918***
	[0.108]	[0.166]	[0.209]	[0.343]	[0.418]	[0.534]	[0.666]	[0.679]	[0.686]	[0.659]
Robust	0.120	0.266	0.301	0.917**	1.284**	0.899	1.236	2.104**	1.719**	1.918**
	[0.131]	[0.198]	[0.243]	[0.409]	[0.507]	[0.668]	[0.812]	[0.829]	[0.842]	[0.805]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	-0.0700	0.0481	0.164	0.433	0.550	0.964	1.478	1.419	1.953	2.376

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Indirect to Agriculture

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010	(10) 2011
Conventional	-0.00565	0.333	0.534	0.519	0.614	-0.0643	0.292	1.637**	1.713***	2.207***
	[0.257]	[0.350]	[0.423]	[0.572]	[0.505]	[0.617]	[0.689]	[0.705]	[0.556]	[0.595]
Bias-corrected	0.0720	0.449	0.670	0.703	0.866*	0.0355	0.357	1.997***	2.075***	2.507***
	[0.257]	[0.350]	[0.423]	[0.572]	[0.505]	[0.617]	[0.689]	[0.705]	[0.556]	[0.595]
Robust	0.0720	0.449	0.670	0.703	0.866	0.0355	0.357	1.997**	2.075***	2.507***
	[0.290]	[0.424]	[0.511]	[0.685]	[0.606]	[0.742]	[0.844]	[0.846]	[0.666]	[0.711]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	0.317	0.199	0.188	0.237	0.257	0.453	0.963	1.039	1.313	1.133

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Percentage change is approximated using difference in logs relative the value reported in 2001. Estimated using local linear regressions with controls for district population and its square. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel.

Table 1.9: Reduced Form RD: Rainfall

Averaged Percentage Deviation from the Mean

VARIABLES	(1) 2000	(2) 2001	(3) 2002	(4) 2003	(5) 2004	(6) 2005	(7) 2006	(8) 2007	(9) 2008	(10) 2009	(11) 2010	(12) 2011
Conventional	-0.0405	-0.0212	-0.0237	-0.0396	-0.00818	0.00883	-0.0638	-0.0373	0.0247	0.0520	0.0633	0.00992
	[0.0796]	[0.0595]	[0.0448]	[0.0444]	[0.0720]	[0.0514]	[0.0707]	[0.0761]	[0.0928]	[0.0707]	[0.0694]	[0.0760]
Bias-corrected	-0.0498	-0.0128	-0.0278	-0.0354	-0.0147	0.00278	-0.0874	-0.0580	0.0359	0.0720	0.0644	0.0131
	[0.0796]	[0.0595]	[0.0448]	[0.0444]	[0.0720]	[0.0514]	[0.0707]	[0.0761]	[0.0928]	[0.0707]	[0.0694]	[0.0760]
Robust	-0.0498	-0.0128	-0.0278	-0.0354	-0.0147	0.00278	-0.0874	-0.0580	0.0359	0.0720	0.0644	0.0131
	[0.0973]	[0.0738]	[0.0532]	[0.0531]	[0.0876]	[0.0624]	[0.0849]	[0.0923]	[0.115]	[0.0872]	[0.0864]	[0.0922]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	95	95	94	94	94	94	94	94	94	94	94
N_UBanked	120	120	120	120	119	120	120	120	120	120	120	120
DepMean	-0.0500	-0.154	-0.0411	-0.208	-0.0298	0.0111	0.0970	0.0506	0.110	0.0545	-0.127	0.174

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Estimated using local linear regressions with no controls. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel.

Table 1.10: RD Results: Individual Crops

Cotton Yield

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010
Conventional	-0.0596	0.192	0.314	0.900	0.444	0.387	0.144	-0.124	-0.380
	[0.205]	[0.358]	[0.521]	[0.681]	[0.563]	[0.632]	[0.649]	[0.676]	[0.595]
Bias-corrected	-0.0693	0.193	0.479	1.165*	0.524	0.596	0.368	-0.128	-0.369
	[0.205]	[0.358]	[0.521]	[0.681]	[0.563]	[0.632]	[0.649]	[0.676]	[0.595]
Robust	-0.0693	0.193	0.479	1.165	0.524	0.596	0.368	-0.128	-0.369
	[0.258]	[0.444]	[0.665]	[0.829]	[0.730]	[0.767]	[0.812]	[0.849]	[0.734]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	55	54	53	37	55	46	51	54	53
N_UBanked	83	83	83	52	72	54	68	69	63
DepMean	1.291	1.327	1.739	1.910	1.771	2.007	1.861	1.757	1.805

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Cotton Output (Bales)

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010
Conventional	15,129	710.2	40,656	62,240	101,178*	176,076**	148,302*	98,366	79,421
	[12,529]	[13,759]	[34,687]	[41,323]	[54,006]	[86,326]	[81,145]	[65,372]	[58,139]
Bias-corrected	15,837	-6,044	53,965	82,530**	113,763**	216,181**	173,677**	110,198*	87,096
	[12,529]	[13,759]	[34,687]	[41,323]	[54,006]	[86,326]	[81,145]	[65,372]	[58,139]
Robust	15,837	-6,044	53,965	82,530	113,763	216,181**	173,677*	110,198	87,096
	[17,640]	[17,434]	[49,786]	[56,638]	[71,802]	[107,247]	[103,780]	[85,300]	[75,475]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	55	54	53	37	55	46	51	54	53
N_UBanked	83	83	83	52	72	54	68	69	63
DepMean	44580	38189	64526	115579	97665	132027	111801	108616	115830

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Wheat Yield

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010
Conventional	0.125	0.124	0.0764	0.173	0.479	0.431	0.459*	0.402	0.313
	[0.143]	[0.167]	[0.166]	[0.158]	[0.426]	[0.291]	[0.240]	[0.260]	[0.216]
Bias-corrected	0.149	0.116	0.126	0.221	0.643	0.546*	0.629***	0.542**	0.369*
	[0.143]	[0.167]	[0.166]	[0.158]	[0.426]	[0.291]	[0.240]	[0.260]	[0.216]
Robust	0.149	0.116	0.126	0.221	0.643	0.546	0.629**	0.542*	0.369
	[0.161]	[0.193]	[0.204]	[0.188]	[0.541]	[0.355]	[0.290]	[0.322]	[0.263]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	58	57	58	59	53	49	55	53	49
N_UBanked	93	88	88	92	88	64	80	87	69
DepMean	2.001	1.944	2.070	1.921	1.969	2.211	2.172	2.212	2.207

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Wheat Output (Tonnes)

VARIABLES	(1) 2002	(2) 2003	(3) 2004	(4) 2005	(5) 2006	(6) 2007	(7) 2008	(8) 2009	(9) 2010
Conventional	13,076	2,897	4,849	6,177	63,023	15,449	49,503	26,082	-10,164
	[11,915]	[10,099]	[18,090]	[20,159]	[56,729]	[30,231]	[39,057]	[36,240]	[27,953]
Bias-corrected	14,898	-1,283	5,066	4,761	79,457	12,576	52,895	27,974	-17,642
	[11,915]	[10,099]	[18,090]	[20,159]	[56,729]	[30,231]	[39,057]	[36,240]	[27,953]
Robust	14,898	-1,283	5,066	4,761	79,457	12,576	52,895	27,974	-17,642
	[14,158]	[12,315]	[21,197]	[22,151]	[73,073]	[36,954]	[46,181]	[46,379]	[35,750]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	58	57	58	59	53	49	55	53	49
N_UBanked	93	88	88	92	88	64	80	87	69
DepMean	161445	146242	162996	151854	158668	196304	191220	204942	194492

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Cotton output measured in bales rather than tonnes. Estimated using local linear regressions with controls for district population and its square, and the pre-randomization 2001 value of the dependent variable. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel.

Table 1.11: RD Results: Crop Yield Index

Reduced Form Separately for Each Year

VARIABLES	(1) 2003	(2) 2004	(3) 2005	(4) 2006	(5) 2007	(6) 2008
Conventional	0.0113	-0.114	0.120	0.380	0.187	0.284*
	[0.101]	[0.146]	[0.132]	[0.268]	[0.157]	[0.170]
Bias-corrected	-0.00324	-0.123	0.117	0.492*	0.226	0.337**
	[0.101]	[0.146]	[0.132]	[0.268]	[0.157]	[0.170]
Robust	-0.00324	-0.123	0.117	0.492	0.226	0.337
	[0.120]	[0.173]	[0.155]	[0.358]	[0.198]	[0.213]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	47	44	38	47	44	45
N_UBanked	74	72	67	74	60	72
DepMean	-0.115	-0.0303	0.0340	0.0133	-0.0710	-0.0449

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Fuzzy RD Instrumenting for Private Bank Credit Accounts, Pre-reform and Post-Reform

VARIABLES	(1) preref	(2) postref
Conventional	-0.0786	0.0290*
	[0.214]	[0.0154]
Bias-corrected	-0.109	0.0362**
	[0.214]	[0.0154]
Robust	-0.109	0.0362*
	[0.247]	[0.0193]
Bandwidth	3.500	3.500
N_Banked	91	174
N_UBanked	146	273
DepMean	-0.0731	-0.0168

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note: Index of crop yield using weighted averages of the crops rice, wheat, jowar and groundnut. Weighted by crop revenue share. Estimated using local linear regressions with controls for district average rainfall percentage deviation from the mean, district population and its mean and the pre-randomization 2002 value of the dependent variable. Bandwidths are set at 3.5 thousand persons per branch and estimated using a triangular kernel. Pre-reform years are considered 2003-2004 and post-reform is 2005-2008.

Table 1.12: Fuzzy RD: Percentage Change in Private Credit Amount to Manufacturing and Processing from 2001 Level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Conventional	-0.141	0.0501	0.0285	-0.367	-0.756	-0.114	0.543	0.985	1.402**	0.913
	[0.528]	[0.614]	[0.533]	[0.488]	[0.624]	[0.590]	[0.631]	[0.610]	[0.655]	[0.686]
Bias-corrected	-0.151	0.139	0.0544	-0.328	-0.860	-0.232	0.703	1.395**	1.824***	1.214*
	[0.528]	[0.614]	[0.533]	[0.488]	[0.624]	[0.590]	[0.631]	[0.610]	[0.655]	[0.686]
Robust	-0.151	0.139	0.0544	-0.328	-0.860	-0.232	0.703	1.395*	1.824**	1.214
	[0.640]	[0.752]	[0.657]	[0.614]	[0.766]	[0.706]	[0.758]	[0.740]	[0.794]	[0.825]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	95	94	94	94	94	94	94	94	94	94
N_UBanked	122	122	121	122	122	122	122	122	122	122
DepMean	0.934	1.098	0.678	0.763	0.553	0.694	1.117	1.231	1.287	1.410

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Percentage change is approximated using difference in logs relative the value reported in 2001. Estimated using local linear regressions with controls for district population and its square. Bandwidths are set 3.5 thousand persons per branch and estimated using a triangular kernel.

Table 1.13: Diff n Diff: States Selected around Under Banked Threshold, 1999-2010

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Ln_Net_Assets	Ln_Working_Capital	Ln_Loans	Ln_Tot_Investment	Cap_Labor_Ratio
TreatXPost2006	0.171	0.264*	0.235*	0.197*	3.426*
	[0.142]	[0.136]	[0.116]	[0.106]	[1.724]
Observations	118,236	101,566	95,269	113,296	118,128
R-squared	0.270	0.195	0.082	0.200	0.012
State FEs	Yes	Yes	Yes	Yes	Yes
State Trend	Yes	Yes	Yes	Yes	Yes

Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1
Standard Errors Clustered at State level

Note: Banked States include Haryana, Uttarakhand, Punjab, Mizoram, Daman and Dimiu, Karnataka and Puducherry. Under Banked States include Rajasthan, Tripura, Jharkhand, Orissa and Dadra and Nagar Haveli. All regressions control for post 2006 and treated state individual effects, logged number of units in firm and the logged number of employees in the enterprise, plant age and its square, a year trend, state specific year trends and state fixed effects.

Table 1.14: Difference in Log Mean District Light from 2004

Discontinuity from Reduced Form Estimated Annually

VARIABLES	(1) 2005	(2) 2006	(3) 2007	(4) 2008	(5) 2009	(6) 2010	(7) 2011
Conventional	-0.0258	0.00485	0.0942***	0.0916***	0.0346	0.0972	0.0810*
	[0.0193]	[0.0272]	[0.0297]	[0.0322]	[0.0707]	[0.0605]	[0.0492]
Bias-corrected	-0.0297	0.00720	0.108***	0.104***	0.0426	0.119**	0.105**
	[0.0193]	[0.0272]	[0.0297]	[0.0322]	[0.0707]	[0.0605]	[0.0492]
Robust	-0.0297	0.00720	0.108***	0.104***	0.0426	0.119*	0.105**
	[0.0210]	[0.0300]	[0.0319]	[0.0353]	[0.0773]	[0.0663]	[0.0531]
Bandwidth	3.500	3.500	3.500	3.500	3.500	3.500	3.500
N_Banked	94	94	94	94	94	94	94
N_UBanked	122	122	122	122	122	122	122
DepMean	-0.139	-0.0808	0.114	0.0722	0.0259	0.355	0.219

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Fuzzy RD Instrumenting for Private Bank Branches, Pre-reform and Post-Reform

VARIABLES	(1) preref	(2) postref
Conventional	-0.0156	0.00482
	[0.100]	[0.00413]
Bias-corrected	-0.0111	0.0119***
	[0.100]	[0.00413]
Robust	-0.0111	0.0119**
	[0.117]	[0.00491]
Bandwidth	3.500	3.500
N_Banked	94	658
N_UBanked	122	854
DepMean	-0.139	0.143

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Note: Reduced form estimated using local quadratic regressions with controls for district population and its square. Bandwidths are set to 3.5 thousand persons per branch and estimated using a triangular kernel. The fuzzy regression discontinuity is estimated using local linear regressions. The number of operating private bank branches is instrumented with predicted under banked assignment. Controls include district population and its square. Pre-reform year is 2005 using 2004 as the base year for the approximate percentage change. Post-reform years are 2006-2012.

Table 1.15: NREGA Discontinuity in District Phase Assignment

	(1)	(2)	(3)
VARIABLES	Phase_1	Phase_2	Phase_3
Conventional	-0.0648	0.0145	0.0503
	[0.119]	[0.0909]	[0.135]
Bias-Corrected	-0.121	0.0710	0.0497
	[0.119]	[0.0909]	[0.135]
Robust	-0.121	0.0710	0.0497
	[0.139]	[0.109]	[0.160]
Bandwidth	3.500	3.500	3.500
N_Banked	93	93	93
N_UBanked	121	121	121
DepMean	0.285	0.201	0.514

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note: Reduced form estimated using local quadratic regressions with controls for district population and its square. Bandwidths are set to 3.5 thousand persons per branch and estimated using a triangular kernel. NREGA was rolled out in 3 phases between 2006 and 2009 based on some measure of expected program need by district.

1.10 Data Appendix

1.10.1 Districts

The majority of analysis in this paper is conducted at the administrative district level in India. Districts constitute the administrative level directly below the state government (and union territory). Data sets at the district level rarely provide numerical identifiers. When available, these identifiers typically do not easily map to other data sets. Further, the anglicized spelling of district names is often inconsistent across and even within data sets. Renaming and redistricting also occur relatively frequently in India. As such, each data set required the assignment of a numerical identifier before conducting analysis. To ensure consistent measures in the data across time, I adjust all data to their 2001 district boundaries from the Population Census. I first assign each district its 2001 state and district numerical codes from the 2001 Population Census, or an auxiliary district code if the district was formed post 2001. Then using the atlas provided in the 2011 Population Census, I map new districts back to their source districts in 2001. Although super-districts, created when newly formed districts drew land from more than one source district, are identified, they are dropped from the analysis.[20] District websites, newspapers and other internet based resources were used to help map alternative spellings to numerical codes.

[20]New districts since 2001 that claimed territory from more than one source district are dropped along with the source districts due to issues with the aggregation. In addition, Thane and Pune districts in Maharashtra are dropped. These districts are located close to Mumbai but are not technically classified as belonging to the greater Mumbai area. They constitute outliers as they achieve rapid growth more likely to be attributable to their proximity to Mumbai. Thane is on the under banked list while Pune is not, though the RBI amended the policy to 2008 to make centers within 100km of Mumbai ineligible for under banked status, effectively removing Thane's status. Varanasi district in Uttar Pradesh is also dropped after 2002 due to the 2002 merger of the private sector Banaras State Bank with Bank of Baroda which is a nationalised bank. Banaras State Bank primarily operated in Uttar Pradesh with the bulk of its branches in districts designated as under banked. However, 20 branches operated in Varanasi which happens to be located right at the cutoff on the banked side. The vast majority of branches affected by the merger belonged to districts designated as under banked. However, the reclassification of 20 branches to public sector bank status just on the banked side of the cutoff results in a sudden drop in the banked intercept in 2003 for private banks. Since most of these branches continued to operate under the public sector, the drop-off creates an exaggerated representation of the policy effect, which does not accurately represent the change to the banking environment. While these branches could be "added back" using the detailed data from the MOF, the same cannot be done for the aggregated data on credit.

1.10.2 Banking

Branches and Licenses Data on the number of operating branches and licenses are from the Master Office File (MOF) accessed from the RBI website in spring 2012. Opening and closing dates (when applicable) are provided for each bank, as well as information regarding branch location and type of business conducted at the branch (e.g. General Banking, Specialized Banking, ATM). "Brick and mortar" branches are used in the analysis, meaning branches classified at being general banking or specialized banking. Not Administratively Independent Offices such as extension counters and ATMs are excluded from analysis. The number of operating branches for each year is calculated as the number of branches with an opening date prior to January 1st of that year and a closing date afterward or missing. Operating branches by subsets of bank group classification are calculated similarly. Licenses are considered to be operating if issued before January 1st of a given year with a branch close date afterward or missing. Thus, licenses can be in "operation" even if branch opening occurs at a later date. After the December 2009 reform granting blanket permissions to open in low population centers, the incidence of unreported license dates for branches in such centers increased. The assumption is made that these constitute branches exercising the blanket permission, such that the effective license date is taken to be the date of branch opening.[21]

Constructing the forcing variable In constructing the forcing variable and national average I follow the APPBO procedure [22] described for identifying deficit districts during the policies of the 1980s and also that for identifying under banked states in the RBI Report of the Group to Review Branch Authorisation Policy (RBI Report, 2009). I take the number

[21] A similar pattern for license dates from branches in urban centers in the Northeast region that had a special exception for blanket permissions for urban centers, and only in that region, provides additional support for this assumption.

[22] The Average Population Per Bank Office was constructed using the district population from the most recent population census, in this case that from 2001, and dividing that by the number of bank offices in that district. I restrict the set of offices to those conducting general and specialized bank business which may depart from the actual algorithm used by the RBI. The national average to which the value is compared is the total population of India divided by the number of bank offices.

of operating branches on September 7th, 2005, the day prior to the 2005 Master Circular issue date that implemented the branching policy reform. Following the rule that Under Banked Status = 1(district population per branch ¿ national average) yields nearly an exact match to the official 2006 list of under banked districts in the 2006 master circular.[23] Out of 572 districts only 6 fail to match their official status. Due to redistricting and the level of aggregation of credit and deposits data, I aggregate all districts bifurcating since 2001 back to their 2001 boundaries. In cases that new districts form from two or more source districts, these are aggregated into a single super district, resulting in 572 districts. Of these, I denote 202 districts as banked (with 204 on the official list) and 370 under banked (368 officially). After dropping super districts from the sample, 4 misassigned districts remain. Replicating the analysis taking the number of operating branches on January 1st, 2006 yields similar results.

Credit The Basic Statistical Returns 1 (BSR1) provides information on credit accounts, credit limits and credit outstanding by scheduled commercial banks including RRBs (last accessed spring 2014). The data are reported annually by banks with values as of March 31st for that year. Credit captured by BSR1 relates to gross bank credit such as term loans, cash credit, overdrafts, etc. Detailed descriptions are provided by the RBI. The financial year 200X-200Y is reported as 200Y in the paper and is reported with consistent notation across analyzed data. Values are delineated by bank group and population group at the district level (e.g. number of credit accounts with Nationalised Banks, by semi-urban areas in Rangareddy). Locations, such as semi-urban Rangareddy, represent the area of credit utilization for loans exceeding 2 lakh Rs. for which detailed account information is collected. Loans of lesser amounts are reported with less information, and the RBI assumes they are utilized in the same area as which the loan was sanctioned. Credit amounts are further

[23] A list of under banked districts was issued with the 2005 master circular. A slightly revised list was reissued with the 2006 master circular and remained unchanged through 2009, after which the districts of some states were dropped. The national average computed using September 7th, 2005 as the policy date was 14,915 persons per branch in India.

delineated by utilization purpose, coined "occupation," and include : agriculture, industry, professional and other services, personal, trade, transport operators, finance and all other. These are broken down further for agriculture into "direct" and "indirect," for industry by "construction" "mining" "manufacturing and processing" and "electricity, gas and water" and trade by "retail" and "wholesale." Personal loans are also presented disaggregated, but the delineation between subgroups appears to be inconsistent through time so are always treated as aggregated personal loans in the analysis. A reclassification of loans to make occupations consistent with a 2004 update of industrial codes occurred in 2008. The reclassification should not have affected aggregate measures of account and amounts, though caution should be taken when attempting to draw comparisons at the occupation level before and after 2008.[24]

The BSR2 provides analogous information for deposits and is structured similarly (last accessed spring 2014). Values are reported for the number of deposit accounts and deposit amounts.

The BSR7 provides quarterly data on credit, deposits and reporting branches. Analysis on BSR7 is not included in this paper.

All credit and deposit limits and amounts are adjusted using the Consumer Price Index for Industrial Workers provided by India's Labour Bureau. I adjust all values to 2011, fourth quarter prices. Amounts are reported in Rupees.

Population Groups The RBI follows a specific assignment procedure for population groups. Based on the Population Census, locations with populations less than ten thousand are designated rural, 10,000 - 100,000 semi-urban, 100,000 - 1 million urban and greater than 1 million metropolitan. Prior to 2005 locations were assigned status based on their 1991

[24] Two districts exhibit measures of credit accounts and amounts that appear to reflect coding errors in the data. Mallapuram, Kerala is dropped in 2004 due to an unexplainable jump in the magnitude of credit unmatched in the district in any other years. Ghaziabad, Gujarat in 2008 displays even more erratic values for certain credit measures. These values are set to missing as the remaining appear unaffected. In both instances, private sector banks with a presence in the concerned district were acquired by the public sector. The reclassification of the bank to the public sector may have created underlying issues in the data reported in those places for those years.

Population Census values. The switch to the 2001 Population Census for reports in 2006 and later make strict comparisons between the sets of years complicated at the disaggregated population group level. The problem appears to be greater for the metropolitan and urban population groups, as fewer centers exist in these categories. The scope for problems appears smaller for rural and semi-urban classifications due to the high volume of centers in these categories. Still, the caveat should be kept in mind for analysis at the disaggregated level.

1.10.3 Agriculture

Crop output and area The data on crop output and area are reported in the Annual Crop Yields at District Level from the Crop Production Statistics. The production output in tonnes and area cultivated in square hectares are reported by crop at the district level either annually or by season, depending on the crop and state. Reported crops vary across districts, and the detail of information on variety and growing season also varies across states and years. I develop the data from a file made available from the Government of India for years 1998-1999 to 2010-2011 (years reported July-June). Extensive cleaning of district and crop names, as well as accounting for redistricting, is required to analyze the data as a panel. I match each district reported to their 2001 Population Census identification number or to a 2011 ID number constructed for this analysis when dealing with new districts since 2001. Analysis is restricted to years 2001-2010 which exhibit lower frequencies of missing data. Missings values after 2010 are reported to be due to unfiled state reports. Districts never reporting positive statistics for a crop over the sample period are dropped from analysis for that individual crop. In years a district reports a missing value for a crop that is reported in that district in other years, the value is interpreted as null and replaced with a zero value.

Crop prices The data on crop prices are from the Farm Harvest Prices of Principle Crops. States are responsible for reporting crop prices for a set of prominent crops each year. The prices are supposed to be collected during the peak harvest times of each crop and account for variations in quality. States vary in their reporting of crop prices by season and detail on variety. Further, states vary in reporting price for some crops by product (e.g. some report prices for sugarcane while others only report prices for raw sugar, cotton lint or whole cotton, etc.) Technical conversion factors for raw crops to agricultural outputs provided by the Statistic Division of the FAO are used where applicable to match prices to corresponding crop outputs. Prices are reported in Rupees per Quintal (an Indian quintal is 100 kg) and must be converted to Rupees per tonne for consistent units with the output data. I have developed the data from pdf reports available in separate sets by state for 2001-2002 to 2003-2004, 2004-2005, 2005-2006, and 2006-2007 to 2007-2008. Efforts to process the remaining years of the data are under way. Extensive cleaning of district names, accounting for redistricting, and assignment to identification numbers was similarly required.

Crop yield index Annual crop yield is calculated as crop output in tonnes per hectare cultivated for that crop. To create the index of crop yields as in Jayachandran (2006), I match the crop prices data to the crop output and area data. Four of the top five revenue producing crops for India identified in Jayachandran (2006) are used in the index, rice, wheat, jowar and groundnut. Sugar is excluded due to concerns regarding the accuracy of conversions of sugarcane to raw sugar production in order to match the two data sets, and whether the reported prices for sugar capture actual prices faced by farmers after accounting for delay of payments bargaining. Crop yields are normalized to have mean values equal to one in each year for comparability across crops. Weighted averages of the log values of the four crop yields are taken for each district year, using the crop revenue share of the total crop revenue of the district from those four crops as weights. When matching the price and production data sets, season and variety matches are made when the detail of data from both sets allow. Otherwise, the mean of price data by district and crop are calculated (if price is broken out by variety or season) and matched to the production data for that crop-year. To increase the number of matches, when prices are missing for a crop at the district level, the weighted state average prices provided in the reports are used. Missing crop prices at the district level generally correspond to relatively low levels of output in the production data. The index is currently constructed for 2002-2008, with efforts to process the remaining years of data under way.

1.10.4 Industry

Annual Survey of Industries The Annual Survey of Industries (ASI) is a detailed survey of registered manufacturing firms in India conducted by the Central Statistical Organisation. The ASI is used extensively in economic research (Hsieh and Klenow, 2009; Bollard et al., 2013) to name just a few). I use fiscal years 2001-2010 in my analysis. In these years, all firms with greater than 100 workers were enumerated, as were all firms operating in the five less developed states/UTs (Manipur, Meghalaya, Nagaland, Tripura and Andaman & Nicobar Islands). The remainder of registered firms (those with greater than 10 workers, assuming compliance) were surveyed from samples representative at the State by NIC-2004 4 digit industry code. In addition to the values reported directly in the ASI, I construct the capital labor ratio as the average of the opening and closing values of assets net of depreciation divided by the sum of the firm's wage bill plus benefits, as in Hsieh and Klenow (2009). Due to the joint census-sampling methodology, I conduct my analysis at the state level in order to apply proper weighting for a representative sample of all registered firms. A thorough discussion of the ASI data can be found in Bollard et al. (2013).

1.10.5 Remote Sensing

DMSP-OLS Nightlights The Defense Meteorological Satellite Program (DMSP) maintains data sets with of night lights data, constituting a yearly average of the amount of light emitted into space at night for a roughly 1km square grid. Using satellite images, algorithms to control for reflection, cloud cover and other confounding factors assign a digital number between 0 and 63 for each cell that may be downloaded as a finely pixelated map of the Earth. Using the boundary outline of India's administrative districts in 2001, I construct the district average of the digital numbers in each district. I then calculate the percentage change of this average as the log of the district mean value minus the log district mean from 2004. Analyzing changes in growth across districts, as opposed to levels is important due to measurement error introduced through machine learning and the algorithms applied to eliminate glare light bleed. I have processed data from satellites F16 and F18, that cover calendar years 2004-2012. Efforts are under way to process the data from F15 that would extend the data set back to year 2000. A thorough discussion of the nightlights data is included in Henderson et al (2012).

TRMM Rainfall Data Rainfall strongly affects agricultural productivity. To the extent that rainfall varies annually across districts, conditioning on it will improve my precision for estimates related to agriculture. I use the publicly available data collected by the Tropical Rainfall Measuring Mission (TRMM) satellite jointly maintained by the National Aeronautics and Space Administration (NASA) and the Japan Aerospace and Exploration Agency (JAXA). Fetzer (2014) gives a detailed description of these data and their verification processes. These data are collected from a satellite orbiting approximately 250 miles above the Earth's surface that completes an orbit several times a day and is able to detect rainfall falling as lightly as 0.7 millimeters per hour. Daily rainfall measures are available from 1998-2012 on a 0.25 by 0.25 degree grid, making it the finest available spatial resolution for India to the best of my knowledge.

These data are likely favorable to those generated using ground rainfall gauges as the latter require local monitoring and maintenance, the quality of which may vary systematically with the prosperity of districts. Further, the spatial diffusion of gauges is not uniform across India, requiring different levels of interpolation between rain gauges that can introduce measurement error that may be difficult to account for and change in less transparent ways as the number and location of gauges vary across time.

1.11 Theoretical Framework Appendix

This section sketches out the theoretical framework for anticipated competition in the second period leading to increased levels of credit at the time of policy announcement in the first. Details of intermediate steps are omitted in the interest of space. The participation constraint for borrower i in each period is given by,

$$\begin{cases} E[\pi_i^1] = P_s(R_s^A)[R_s^A(1-\theta_i) - (1+r^1)] - s > \mu & Period\,1 \\ E[\pi_i^2] = \{P_s(R_s^A)[R_s^A(1-\theta_i) - (1+r^2)] - s1(First\,Time\,Borrower) > \mu & Period\,2 \end{cases} \tag{1.3}$$

where $\theta_i \sim uniform[0,1]$ is a privately known cost to the borrower that is constant across periods (as is being a safe type), r^t denotes the interest rate in period 1 or 2, and $1(FirstTimeBorrower)$ is the indicator function for the borrower's first period of borrowing from the specific bank. Consider the borrower participation constraint from period t. The indifferent borrower with type θ_i facing interest rate r^t will satisfy

$$P_s(R_s^A)R_s^A(1-\theta_i) = P_s(R_s^A)(1+r^t)] + s1(First\,Time\,Borrower) + \mu \tag{1.4}$$

Rearranging terms, the indifferent borrower may be expressed as a function of the interest rate r^t,

$$\hat{\theta}_i(r^t) = 1 - \frac{P_s(R_s^A)(1+r^t) + s1(First\,Time\,Borrower) + \mu}{P_s(R_s^A)R_s^A} \tag{1.5}$$

such that all borrowers with $\theta_i < \hat{\theta}_i(r^t)$ will demand a loan with interest rate r^t. For a market of size M, total demand for loans at interest rate r^t will be $M\beta\hat{\theta}_i(r^t)$.

Assume banks are profit maximizers, face an exogenous marginal cost of funds plus administrative costs of lending equal to $(1+\rho)$, and cannot discriminate in the interest rate offered to repeat and first time borrowers. The bank's participation constraint from each period is,

$$E[\pi_B^t] = P_s(R_s^A)(1+r^t)\theta(r^t) > (1+\rho)\theta(r^t) \quad for\, t = 1,2 \tag{1.6}$$

In deciding the interest rate for each period, the incumbent bank will anticipate its outcome in the second period if facing entry and take that into consideration in setting its first period interest rate. Specifically, if a new bank enters the market in the second period, the incumbent will expect to compete in interest rates such that the entrant must offer his zero-profit condition interest rate and the incumbent will offer the interest rate making his first period borrowers that do not pay the screening fee if they stay indifferent between borrowing from him and the incumbent.

Sketch of Proof: *If the incumbent offers an interest rate higher than that making first period borrowers indifferent between borrowing from the incumbent while avoiding switching costs and borrowing from the entrant while incurring the switching costs, then the incumbent loses the entire market to the entrant. If the incumbent offered an interest rate lower than that value, then he loses profits from the locked in first period borrowers but gains no new borrowers since new borrowers must pay the screening fee regardless and the entrant's interest rate is strictly lower. If instead the entrant offered a price above the zero profit condition interest rate, then the incumbent would increase his rate to earn higher profits off of his first period borrowers. This, however, creates incentive for the entrant to lower his interest rate a small amount and capture the entire market. If the entrant instead lowers his interest rate he will serve the entire market at a loss.*

Taking the second period equilibrium into consideration, the incumbent knows his second period interest rate when facing entry will be $1 + r_2^I = \frac{1+\rho+s}{P_s(R_s^A)}$ by equating demand for the zero profit interest rate and demand for an interest rate when the switching cost need not be incurred. The incumbent will then maximize first period interest taking the second period predetermined interest rate into consideration as the first period interest rate will determine the demand faced in both periods. Thus, the incumbent's maximization problem is

$$\max_{r_1^I, r_2^I} P_s(R_s^A)(1+r_1^I)\theta(r_1^I) + \delta P_s(R_s^A)(1+r_2^I)\theta(r_2^I) - [(1+\rho)\theta(r_1^I) + \delta(1+\rho)\theta(r_2^I)] \quad (1.7)$$

Substituting in the value for r_2^I and setting demand equal in both periods reduces the problem to

$$\max_{r_1^I} P_s(R_s^A)(1+r_1^I)\theta(r_1^I) + \delta P_s(R_s^A)(\frac{1+\rho+s}{P_s(R_s^A)})\theta(r_1^I) - (1+\delta)(1+\rho)\theta(r_1^I) \quad (1.8)$$

Taking the first order condition with respect to r_1^I, setting it equal to zero and solving for the optimal first period interest rate for the incumbent yields,

$$1 + r_1^{I*Entry} = \frac{1}{2P_s(R_s^A)}\{P_s(R_s^A)R_s^A - (1+\delta)s - \mu + (1+\rho)\} \quad (1.9)$$

Intuitively, the incumbent increases the interest rate with the expected payoff of the project to capture additional surplus as well as the cost of lending the funds and lowers the interest rate with the borrower's reservation utility. The incumbent lowers the interest rate as the switching cost increases, as this relaxes the constraint on the interest rate he offers in the second period, allowing for higher profits from each continuing first period borrower.

To determine the effect of anticipated competition on first period lending, consider an incumbent that does not expect entry in the second period. He will find it optimal to set interest rates so as to maximize total profit from both periods, increasing the interest rate in the second period to extract the additional surplus from the repeat borrowers no longer paying the screening cost. Since no other changes occur to the environment, the incumbent will maximize profits by serving the same set of borrowers in both periods, setting the second period interest rate so as to make the marginal borrower indifferent between accepting the loan and not. The maximization for the incumbent not expecting entry may be expressed as,

$$\max_{r_1^I} P_s(R_s^A)(1 + r_1^I)\theta(r_1^I) + \delta P_s(R_s^A)(1 + r_1^I + s)\theta(r_1^I) - (1 + \delta)(1 + \rho)\theta(r_1^I) \tag{1.10}$$

Taking the first order condition with respect to r_1^I, setting it equal to zero and solving for the optimal first period interest rate for the incumbent yields,

$$1 + r_1^{I*NoEntry} = \frac{1}{2P_s(R_s^A)}\{P_s(R_s^A)R_s^A - (1 + \frac{\delta P(R_s^A)}{(1 + \delta)})s - \mu + (1 + \rho)\} \tag{1.11}$$

Finally, since the interest rate determines the first period quantity of credit, anticipated competition will lead to an expansion of credit if $1 + r_1^{I*Entry} < 1 + r_1^{I*NoEntry}$. This inequality reduces to the simple expression, $\frac{P_s(R_s^A)}{1+\delta} < 1$ that must always be true. Hence, introducing the potential of future competition leads to an expansion of credit at the time announcement.

Chapter 2

Commercial Banking and Consumption Smoothing

2.1 Introduction

Exposure to risk can have dramatic effects on consumption and productivity. Households in developing countries, particularly in rural areas, can face vast variability in incomes relative to earners in the United States (Rosenzweig and Binswanger, 1993). Institutions, formal and informal, arise to help individuals cope with these risks. Townsend (1994) examines the ability of households to insure against income shocks, and finds close to full insurance within the ICRISAT villages of India. The lack of adequate financial markets, as shown in Rosenzweig and Wolpin (1993), can lead households to hold buffer stocks and make asset adjustments to help smooth consumption across periods of fluctuating income. The informal institutions frequently available in rural areas (money lenders, family members, friends, etc.) can be ill-suited to insure against systematic risks shared across the local market. Realized rainfall, being an important agricultural input and spatially correlated, constitutes just such a shock. Rosenzweig and Binswanger (1993) show that total farm profits are significantly affected by rainfall. Consistent with incomplete markets for instruments to smooth consumption following realized income, they find farmers adjust their investment decisions toward less rain-sensitive crops that also exhibit lower mean returns. While this strategy reduces fluctuations, it also leads farmers to operate at lower levels of efficiency. Further, households at the lower end of the wealth distribution, that may be more likely to lack access to financial markets and suffer worse outcomes from negative shocks, are shown to adopt this form of insurance more. Rainfall risk therefore not only leads to lower levels of efficiency, but can contribute to growing income inequality.

Evidence from recent work suggests that rainfall risk continues to distort production activities in India. In a randomized control trial setting, Cole et al. (2013b) show that farmers adopting rainfall insurance shift their investments toward agricultural inputs that are more rain sensitive while generating higher mean returns. Similarly, Rosenzweig and Udry (2013) show rainfall uncertainty leads farmers to under-invest relative to predicted optimal levels, and that farmers inputs are very sensitive to monsoon forecasts, particularly in areas where the forecasts exhibit greater accuracy. Those farmers responding to monsoon forecasts also exhibit greater variances in income, due to the lower performance of their crops when forecasts are incorrect. While rainfall insurance provides an ex ante solution to rainfall uncertainty, Mobarak and Rosenzweig (2014) show that insurance targeted exclusively to landowners can lead to negative effects for the landless in times of low rainfall realizations, exacerbating income volatility for individuals already at-risk. Lower than expected take-up rates of the insurance also diminish its effectiveness against risk. Cole et al. (2013a) find evidence that rainfall insurance faces steep demand curves, but that non-price factors such as distrust, a lack of understanding of the instrument and liquidity constraints are important in hindering adoption.

In this chapter, I provide evidence that the adjustment of deposits and credit from private sector banks in India is consistent with the provision of consumption smoothing with respect to variation in rainfall. The response of investment decisions to anticipated rainfall and insurance in the literature indicates a lack of complete markets to smooth consumption. The expansion of India's private sector banks since deregulations in the early 1990s and branching policy reforms since 2005 makes it an important contributor toward completing financial markets. I examine district level banking data on deposits, credit and branches from India's central bank and rainfall data collected by remote sensing for the period 2001-2011. Consistent with households utilizing banks to smooth consumption, I find deposits increase following periods of good rainfall, while credit increases following lower realized rainfall.

In the next section, I describe the empirical strategy, followed by a description of the data. In section 4 I discuss results and close with a discussion of the results and the next steps for the analysis in section 5.

2.2 Empirical Strategy

I use within-district variations in rainfall to identify periods of relatively higher and lower productivity. Rainfall in India has been shown correlate positively with farmer profits (Rosenzweig and Binswanger, 1993). Changes in productivity are hypothesized to shift the local demand and supply of funds. Agriculture, which constitutes an important sector of the Indian economy, will be more productive when experiencing high rainfall. Periods of higher rainfall are expected to be periods when households will have higher disposable income due to good agricultural conditions. In these periods they will demand more deposit accounts and wish to save more. During periods of low rainfall, households will be expected to dissave, using some of their buffer stock savings from deposit accounts to smooth consumption. If households are utilizing saving deposits in banks to support consumption smoothing, deposit measures are expected to positively correlate with rainfall.

The response from credit to changes in agricultural productivity is theoretically ambiguous. High productivity states may result in greater amounts of loans being made if banks view projects as being less risky or offering higher returns in these periods. Alternatively, the demand for loans may be highest during low productivity periods as farmers substitute towards more expensive inputs. The need for consumption loans during this period is also likely to be higher. Thus, the correlation between credit and rainfall is an empirical question. However, from the relationships proposed here, a negative correlation will be more likely to support consumption smoothing and loans for substitute inputs, while a positive sign either indicates a lack of consumption loans being made or at the least its dominance by productive and "safer" loans.

I estimate a fixed effects model of banking statistics on measures of rainfall including district and year fixed effects to account for unobservable but time-invariant district traits that may effect banking and other outcomes. I estimate the equation,

$$\begin{aligned} y_{it} = \quad & \alpha + \beta_1 log(rainfall_{it}) + \beta_2 log(rainfall_{it-1}) \\ & + \beta_3 log(branches_{it}) + \delta^i d_i + \delta^t d_t + \epsilon_{it} \end{aligned} \tag{2.1}$$

where y_{it} denotes a logged banking outcome such as credit amounts at the district i and year t levels. The effects of rainfall on deposit amounts and accounts, credit limits, amounts and accounts and $log(\frac{Credit}{Deposit})$, will be estimated. The log of rainfall and its lag are both included to flexibly account for growing patterns and translating agricultural outcomes into financial sector effects. The coefficients of interest will be β_1 and β_2, that are estimates of the elasticities of outcomes with the history of rainfall.

2.3 Data

Data for the analysis are drawn from multiple sources. Deposit and credit level data held by scheduled commercial banks are from the Basic Statistical Returns (BSR) II and I, respectively, maintained by India's central bank, the Reserve Bank of India (RBI). Banks annually report the number of deposit accounts and amounts by branch to the RBI with information on account, bank and branch characteristics including location. The RBI makes district level aggregates publicly available, delineated by bank group (Nationalised Banks, Private Sector Banks, Regional Rural Banks, etc.). Banks similarly report on the credit they have extended[1]. The number of accounts, credit limits and actual credit amount lent are reported by their area of *utilization*[2]. Digitized reports are available from 2001-2011. Though branch level data by area of utilization would be ideal, the district level values are more geographically specific than credit data reported by banks in Call Reports in the United States.

Detailed information on the opening, closing and location of scheduled commercial bank branches are available from the Master Office File (MOF) of bank offices maintained by the

[1] BSR I and II report values at the end of the Indian fiscal year, March 31st.

[2] Information on accounts with credit limits below a certain amount are bundled together for reporting and assumed to be lent locally with respect to the reporting branch.

RBI. Banks must register all branches with the RBI, as well as subsequent changes in branch location and closures. From the MOF, a detailed data set may be constructed of bank specific branch networks. The analysis considers brick-and-mortar branches, excluding ATMs, extension windows, satellite offices and other enterprises that do not maintain a separately reported set of books.

The rainfall data used in this analysis are from the publicly available data collected by the Tropical Rainfall Measuring Mission (TRMM) satellite jointly maintained by the National Aeronautics and Space Administration (NASA) and the Japan Aerospace and Exploration Agency (JAXA). Fetzer (2014) gives a detailed description of these data and their verification processes. These data are collected from a satellite orbiting approximately 250 miles above the Earth's surface that completes an orbit several times a day and is able to detect rainfall falling as lightly as 0.7 millimeters per hour. Daily rainfall measures are available from 1998-2012 on a 0.25 by 0.25 degree grid, making it the finest available spatial resolution for India to the best of my knowledge. The data used here are the monthly averaged reported data on rainfall in millimeters. I sum rainfall for each geographic location to the amount falling in the year, April 1st to March 31st. I define the annual district rainfall as the average of total rainfall for geographic locations with coordinates falling within 2001 district borders.

2.4 Results

Table 2.1 presents the results from the fixed effects model of banking statistics on rainfall. In each regression the log of the banking statistic is regressed on the log of district rainfall and its lag, with a full set of district and year fixed effects. Since the purpose of these regressions is to estimate the sensitivity of banking measures to rainfall, observations in which the measures are not reported due to a lack of an operating branch are dropped. As discussed above, bank selection of districts in which to branch according to their elasticities of demand for banking services with rainfall may result in bias. Future work will seek to address these concerns by modeling that selection.

Column 1 presents the results of the regression on logged deposit amounts. Deposits do not show a statistically significant elasticity in rainfall levels for the year preceding March 31st when bank data are reported, though the sign is positive. In contrast, lagged rainfall exhibits a 0.115 elasticity significant at the 5% level, such that a 1% increase in rainfall is estimated to produce a 0.12% increase in deposits. The effect of logged operating branches is reasonable at 1.38 and statistically significant at the 1% level. The elasticity greater than one suggests the potential for competition effects between branches. Deposit accounts produce similar results. The coefficient on operating branches is closer to 1, consistent with diseconomies of scale found in Aguirregabiria et al. (2014). Branches may be capacity constrained in managing or perhaps soliciting deposits, such that more branches are required in order to expand deposits.

Looking to credit reported in column 3, contemporaneous rainfall continues to exude little effect. Lagged rainfall instead negatively effects credit limits. A 1% increase in rainfall is estimated to reduce the total credit limit for the district by 0.22%, significant at the 5% level. The effect of new branches on credit is smaller than it was for deposits at 0.82 and precisely estimated. This could be consistent with a finite supply of investment-worthy projects in an area, or at least projects with reasonably verifiable quality. Credit amounts yield similar outcomes. The estimated effects of rainfall and branches on the number of credit accounts display similar sign patterns, but generally lower magnitudes. This evidence is consistent with the bank exerting control over the issuance of new loans during times of economic shock, but less able to dampen the amount of credit demanded by customers already with a credit line, who were perhaps previously operating below their limit.

Finally, column 6 offers a measure of how credit moves relative to deposits within a district relative to changes in rainfall. Rainfall and its lag both affect the log of credit divided by deposit amounts negatively, with an elasticity of -.27 for lagged rainfall significant at the 1% level. The decrease with rainfall suggests a reduction in credit, at least relative to deposits, or an increase in deposits, or both.

2.4.1 Robustness Check

In the appendix to this chapter, I provide similar results for the Nationalised Banks sector in India. While deposit levels respond to rainfall with statistical significance, no effects on credit are observed. As a robustness check that the results from the private sector are not driven by selection, as they operate in fewer districts than the set of nationalised banks, I perform the analysis for nationalised banks including only those districts served by private sector banks. The results are qualitatively and quantitatively similar, suggesting the effects observed for the private sector are not driven by selection.

2.5 Conclusions

The response of credit and deposit levels of private sector banks to rainfall is consistent with the banks providing consumption smoothing instruments for households with respect to variations in income driven by realized rainfall. In addition to reducing the volatility of consumption, the literature argues there may be gains to agricultural efficiency as farmers adjust their investments and crop selections. Future work should investigate potential shifts in crop composition with the expansion of private sector banks into new districts, as well as explore the potential that banks take the covariance of rainfall patterns when making decisions with respect to their network of branches.

2.6 Tables

Table 2.1: Private Sector Banking Response to Rainfall

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Deposit Amounts	Deposit Accounts	Credit Limit	Credit Amount	Credit Accounts	Log(Credit/Deposits Amounts)
Log(Rainfall)	0.0439	0.0530	0.0125	-0.0220	0.0436	-0.0921
	[0.0543]	[0.0415]	[0.0890]	[0.0831]	[0.0680]	[0.0918]
Lag Log(Rainfall)	0.115**	0.130***	-0.221**	-0.212**	0.0599	-0.269***
	[0.0529]	[0.0421]	[0.102]	[0.0916]	[0.0748]	[0.0911]
Log(Private branches)	1.376***	1.057***	0.820***	0.812***	0.573***	-0.611***
	[0.117]	[0.0609]	[0.117]	[0.108]	[0.0615]	[0.179]
Constant	3.444***	0.569	6.193***	6.588***	-0.254	2.312*
	[0.699]	[0.521]	[1.056]	[1.034]	[0.902]	[1.175]
Observations	2,015	1,987	2,338	2,331	2,338	2,008
R-squared	0.919	0.954	0.852	0.858	0.921	0.594
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Bandwidth	4	4	4	4	4	4

Robust standard errors in brackets
*** $p<0.01$, ** $p<0.05$, * $p<0.1$
Standard Errors Clustered at District Level

Note: The dependent variable is the logged value in each specification. Banking statistics from the BSR 1 and 2. Excludes observations where the banking value was not reported.

2.7 Appendix

Table 2.2: Nationalised Sector Banking Response to Rainfall

VARIABLES	(1) Deposit Amounts	(2) Deposit Accounts	(3) Credit Limit	(4) Credit Amount	(5) Credit Accounts	(6) Log(Credit/Deposits Amounts)
Log(Rainfall)	-0.0215	0.0330***	0.0234	0.0341	0.0189	0.0413
	[0.0226]	[0.0121]	[0.0441]	[0.0444]	[0.0156]	[0.0470]
Lag Log(Rainfall)	-0.0123	0.0244*	0.0162	0.0255	0.0124	0.0420
	[0.0201]	[0.0124]	[0.0559]	[0.0499]	[0.0164]	[0.0462]
Log(National branches)	0.576**	0.335***	0.843**	0.886***	0.137*	0.229
	[0.227]	[0.0923]	[0.374]	[0.337]	[0.0812]	[0.219]
Constant	3.901***	0.487***	1.525**	1.247**	-0.0211	-2.278***
	[0.263]	[0.175]	[0.650]	[0.614]	[0.209]	[0.587]
Observations	2,534	2,533	2,618	2,618	2,618	2,534
R-squared	0.991	0.994	0.951	0.958	0.990	0.752
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Bandwidth	4	4	4	4	4	4

Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1
Standard Errors Clustered at District Level

Note: The dependent variable is the logged value in each specification. Banking statistics from the BSR 1 and 2. Excludes observations where the banking value was not reported.

Table 2.3: Nationalised Sector Banking Response to Rainfall: Restricted Set of Districts

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Deposit Amounts	Deposit Accounts	Credit Limit	Credit Amount	Credit Accounts	Log(Credit/Deposits Amounts)
Log(Rainfall)	-0.00374	0.0360**	0.0262	0.0451	-0.00607	0.0413
	[0.0244]	[0.0142]	[0.0465]	[0.0436]	[0.0163]	[0.0470]
Lag Log(Rainfall)	0.00913	0.0317**	0.0334	0.0430	-0.00648	0.0420
	[0.0205]	[0.0139]	[0.0531]	[0.0459]	[0.0152]	[0.0462]
Log(National branches)	1.330*	0.429***	0.894**	0.930***	0.109	0.229
	[0.744]	[0.124]	[0.368]	[0.330]	[0.0816]	[0.219]
Constant	1.020***	-0.357**	0.473	0.964	-0.110	-2.278***
	[0.210]	[0.180]	[0.606]	[0.607]	[0.194]	[0.587]
Observations	2,015	1,987	2,338	2,331	2,338	2,534
R-squared	0.988	0.994	0.958	0.966	0.989	0.752
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Bandwidth	4	4	4	4	4	4

Robust standard errors in brackets
*** $p<0.01$, ** $p<0.05$, * $p<0.1$
Standard Errors Clustered at District Level

Note: Analysis includes only those districts where private sector banks reported a positive value for that banking statistic.

Chapter 3

Estimating a Sequential Search Model with Bayesian Learning: A Case of Online Search for Differentiated Goods (with Sergei Koulayev)

3.1 Introduction

The advent of online shopping has brought the notion of conducting a search for a product to the commonplace. However, much about the strategies and processes employed by agents during searches remain unknown. Traditional approaches to search have relied on an assumption that agents know the distribution from which they are searching in order to form beliefs regarding the value of search (Stigler, 1961; McCall, 1970; Mortensen, 1970). De los Santos et al. (2012) show that models assuming such known distributions cannot explain certain observed behaviors from observed search, such as the return to previously viewed alternatives. To explain such recall, reservation utilities must adjust during search. Models that exhibit increasing costs to search, or that allow agents to learn about the distribution of potential alternatives can generate these changes (Koulayev, 2014). In this chapter, a coauthor and I relax the assumption that agents know the distribution from which they are searching. We estimate a model of sequential search for a differentiated product, hotel rooms shopped for on a popular search site, introducing Bayesian learning over the distribution of prices as agents form their beliefs.

Rothschild (1974) argues the assumption of known distributions of alternatives from search is unrealistic, and proposes a model where agents must learn about even the distribution from which products are drawn. Agents learn through product draws and updating Dirichlet priors. Koulayev (2013) applies Dirichlet priors to introduce learning into a search model that may be estimated from market share data. In a concurrent and independently developed paper to this chapter, De los Santos et al. (2013) relaxes the assumption of the known distribution. Similar to this chapter, they estimate a search model for differentiated goods where beliefs about future draws adjust through learning. They model agents as knowing potential utility values, but being unaware of the distribution from which they are drawn. They update their belief over the probability of observing any particular utility during sampling using Dirichlet process priors according to Bayes rule. A consequence of modeling learning using the Dirichlet process priors is that reservation utilities necessarily decline with search. This mechanically eases the burden placed on search costs to explain discontinued search, resulting in lower estimated search costs. In this chapter, we will instead assume that agents do not know the distribution of prices, but do know that they are drawn from a normal distribution. Learning modeled in this way offers the benefit of flexibility in forming beliefs, with reservation utilities rising and falling as agents become more optimistic or pessimistic regarding the potential payoffs from search. Changes in agent-specific beliefs occur based on their sampled price draws relative to their prior expectations. The relative distance of draws from expectations will also yield varying effects on beliefs, unlike the Dirichlet priors. Modeling learning with these extra flexibilities will help disentangle the effects of agent learning from their individual draws of search costs, improving the estimation of the parameters defining the search cost distribution. Separating learning effects from search costs may assist in choosing between policy recommendations that target one effect or the other.

We consider a two period model for simplicity. Since only one decision node exists prior to the terminal node, beliefs are only constructed once and expectations over second page results must only be computed once, yet we are still able to allow return to previous results. If the model were extended to multiple period search, the model would support the recall of earlier alternatives before exhausting the full search set. The sequence of agent search actions are observed in the data, as are the characteristics of the products in the agent choice sets, and the final product selected. Placing structural assumptions on the distribution of search costs and utility shocks, we are able to identify the parameters of taste for observed characteristics and the parameters defining the distribution of search costs. We estimate the parameters of the search cost distribution.

We find a slightly lower median search cost in the model with Bayesian learning as compared to the model forming beliefs from the empirical distribution. Using the estimated coefficients on hotel characteristics and estimated parameters for the search cost distribution, we conduct counterfactuals to estimate the change in demand resulting from placing the most popular hotel choices on the first page of results. We show that placing the most popular options on the first page induces an increase in the number of agents selecting hotels.

Methodologically, the learning process modeled in this chapter is similar to the work on learning found in Ackerberg (2003) and Crawford and Shum (2005). Both papers exploit natural conjugate priors of the normal distribution similar to this analysis. The work is related to other analyses estimating search costs that instead apply the empirical distribution (Hortaçsu and Syverson, 2004; Hong and Shum, 2006). As the variety of methods in the above mentioned literature demonstrate, there is not an agreed upon model for the learning process. Salmon (2001) and Nyarko and Schotter (2002) examine the methods of belief updating in experiment settings. In future work, other modes of learning should be considered and evaluated in comparison.

There are a few important dimensions along which our analysis departs several of those stated above. The learning in Ackerberg (2003) is in the agent specific mean utility of consumption which remains constant across time. In our model, agents will be updating their beliefs in regard to mean prices in the market, with mean prices varying for each page of results. Another key difference is the multitude of products and relatively large number of characteristics considered in our analysis. The next section will discuss the details of how the search model developed in Koulayev (2014), into which our model of learning is embedded, address the heterogeneity of products.

3.2 Model

Preliminaries

Our product and search assumptions are intended to capture the important traits of the actual internet search environment experienced by the agents. Agents enter the market by submitting a request to see rooms for a city that are available over certain dates and can accommodate a set number of occupants. An algorithm from the site then generates an initial page of hotel options. Agents view a set of fifteen hotel options on each page they visit. After viewing a page of results, the agent must choose whether to select one of their previously observed results, including the outside option to leave the site without making a selection, or to engage in search to uncover the next page of results. We assume that agents view at most two pages of results in this simple model, that search requires incurring a cost, and that agents update their beliefs over the distribution of hotel characteristics in a Bayesian manner based on the options they observe on the first page.

Certain restrictions on search behavior, as well as assumptions on agent psychology, must be asserted to validate the model framework. First, we restrict our attention to those agents pursuing default search results, i.e. agents do not use filters for characteristics such as hotel star rating or sort by prices, etc. Allowing for the use of filters and sorting introduces many difficulties into modeling and estimation that do not directly contribute to understanding the process of learning and its relationship to search costs. Agents who apply filters and sort results may be drawn from a different population, or populations, than default searchers, which is an issue that lies beyond the scope of this paper. In recent work, Chen and Yao (2014) addresses the use of filters in sequential work. We also assume there exists perfect recall of results, and costless return, so that agents view all options as equally viable regardless of their position within search. Finally, we assert that agents do not exhibit a "first-page bias" which may result from a belief that the website populates search results in a decreasing fashion beginning with its best guess as to a match for the agent, or from imperfect recall that makes first page return easier than returning to middle pages. This last assumption may be quite strong, as the first page is the majority choice given any length of search, conditional on return, and the second most popular choice overall, after the outside option.

3.2.1 The Decision to Search

Consider an agent who views one page of results. Fifteen options with varying characteristics including price are available, plus the outside option. The agent can identify the option out of the fifteen that yields the highest utility for them; denote this option u_{1i}, where the subscript 1 indicates the page the option is drawn from and i is the consumer index. The agent also knows their utility level of the outside option, denoted u_{0i}. Let us denote the maximum of the observed first page options with the notation

$$u_{1i}^{*} = max\{u_{0i}, u_{1i}\} \tag{3.1}$$

The utility u_{1i}^{*} indicates the utility level attainable by the agent without engaging in search. Once observing u_{1i}^{*} the agent must decide whether or not to engage in search. We assume the agent weighs her expected benefit from getting to choose from the second page of results against the cost incurred to view that page. If she views the second page, she will choose the observed option that yields the highest utility, $u_{2i}^{*} = max\{u_{0i}, u_{1i}, u_{2i}\}$ which may be written iteratively as $u_{2i}^{*} = max\{u_{1i}^{*}, u_{2i}\}$. Note, the agent cannot know u_{2i}^{*} when only observing the first page results, prior to engaging in search. A simple implication of this process is that $u_{2i}^{*} \geqslant u_{1i}^{*}$ since search merely expands the choice set, and the utility from any option does not reflect the cost involved in finding it. If a better option is not found on the second page, the agent may still acquire the utility from the best option prior to search.

3.2.1.1 The Value of Search

We assume each agent draws an individual search cost denoted, c_i. Then the option value from observing the first page can be characterized in the following value function,

$$V_1(u_{1i}^{*}, c_i) = max\{u_{1i}^{*},\, E_1(V_2(u_{2i}^{*})|\Omega_{1i}) - c_i\} \tag{3.2}$$

where Ω_{1i} denotes the information set available to the agent after observing the first page of results.

The second period is the terminal period, so the value of being at that page is simply the utility from the best option,

$$V_2(u_{2i}) = u^*_{2i} \tag{3.3}$$

Note that the value function in the terminal period does not depend on search cost, as previous search costs are sunk, and no future search is available. Substituting in the expression for the second period value function in equation (3), we can define the expected value function for the first page,

$$Q_1(u^*_{1i}) \equiv E_1(u^*_{2i}|\Omega_{1i}) \tag{3.4}$$

which expresses the expected value of results on the second page, without considering the cost required to view them.

3.2.1.2 Reservation Utilities

The decision to search in the first period then comes down to a tradeoff between the expected value function and the cost of search relative to the best option from the first page. Agents will be indifferent between searching and keeping their current highest utility option when

$$Q(u^*_{1i}) - c_i = u^*_{1i} \tag{3.5}$$

Rearranging terms and substituting in the expression for the expected value function, we can consider instead $E_1 max(u_{2i}, u^*_{1i}|u^*_{1i}) - u^*_{1i} = c_i$. The left hand side of this equation we argue is monotonically decreasing in its argument u^*_{1i}. As the baseline utility from the first page increases, the potential gain from search will be relatively lower. At each u^*_{1i} level, not every second page draw is expected to be an improvement in quality. As u^*_{1i}

increases, the probability that the draws from the next page will produce an improvement diminishes for reasonable distributions. Thus, $Q(u_{1i}^*)$ will increase at a decreasing rate, and $\triangle u_{1i}^* < Q(\triangle u_{1i}^*)$, satisfying our claim that the left hand side of the rearranged equation is decreasing in u_{1i}^*.

With $Q(u_{1i}^*) - u_{1i}^*$ monotonically decreasing, and c_i fixed at a constant, there exists a value of u_{1i}^* for which the equality holds. Let us denote that value r_{1i} to represent the reserve utility for the agent. Thus, an agent will engage in search whenever $u_{1i}^* < r_{1i}$, and stop searching when the inequality is violated.

3.2.1.3 Revealed Preferences Interpretation

We will follow a revealed preference interpretation of search decisions and option selections, alternatively referred to as "clicking". As such, the length of search and the position of the selected option across pages have important implications on the utility relationships of the hotel options and reservation utilities. These inferred relationships are summarized in the table below, where k denotes the page of the selected hotel, and t denotes the number of pages searched.

	$t = 1$	$t = 2$
$k = 0$	$r_{1i} < u_{0i}$ $u_{1i} < u_{01}$ ambiguous: $u_{1i}?r_{1i}$	$u_{0i},\ u_{1i} < r_{1i}$ $u_{2i} < u_{0i}$ $u_{1i} < u_{0i}$
$k = 1$	$r_{1i} < u_{1i}$ $u_{1i} < u_{0i}$ ambiguous: $u_{01}?r_{1i}$	$u_{0i},\ u_{1i} < r_{1i}$ $u_{2i} < u_{1i}$ $u_{0i} < u_{1i}$
$k = 2$	-	$u_{0i},\ u_{1i} < r_{1i}$ $u_{0i} < u_{2i}$ $u_{1i} < u_{2i}$

We will leverage these inequalities when taking the probabilities of joint events to use in our estimation strategy.

3.2.2 Utility

We model the utility from a particular hotel as a linear function in the price and characteristics of the hotel and a random utility shock drawn from a standard Extreme Value Type 1 distribution. The utility shock may be interpreted as an individual specific match value. We can represent the parameters and characteristics in vector form, $\theta = (\alpha_{do}, \alpha_d, \alpha_s, \overrightarrow{\alpha_n}, \overrightarrow{\alpha_b}, \alpha_P)$ and $q_{ij} = (do_j, d_j, s_j, \overrightarrow{n_j}, \overrightarrow{b_j}, P_{ij})$, respectively. The vector of characteristics is individual specific because the observed price, P_{ij}, can vary across agents for the same room j. The set of characteristics considered is distance to O'Hare Airport, distance to the city center, the star-rating, a set of indicators for the neighborhood, a set of indicators for the hotel chain, and the price of the room, respectively. Thus, the utility function may be written,

$$u(p_j, q_{ij}, \varepsilon_{ij}) = \alpha_{do} do_j + \alpha_d d_j + \alpha_s s_j + \overrightarrow{\alpha_n}\overrightarrow{n_j} + \overrightarrow{\alpha_b}\overrightarrow{b_j} + \alpha_P P_{ij} + \varepsilon_{ij} \tag{3.6}$$

and expressed as the sum of a mean utility and taste shock,

$$u(p_j, q_j, \varepsilon_{ij}) = \mu_{ij} + \varepsilon_{ij} \tag{3.7}$$

where $\mu_{ij} = \theta q_{ij}$ in our base line model where consumers exhibit homogenous tastes for characteristics.

Many initial requests result in leaving the site without clicking a single option. The consumer in this case is considered to have selected the outside option. The utility for the outside option is denoted, $u_{i0} = \mu_{out} + \overrightarrow{\mu_0}\overrightarrow{R_i} + \varepsilon_{i0}$ where $\overrightarrow{R_i}$ denotes a vector of indicator variables for the request parameters submitted by the agent to instigate the first page.

3.2.3 Beliefs

To compute the reservation utilities, agents must take an expectation with respect to the maximal utility from the yet unobserved next period results to assess the benefit of search. To do so, the agent must formulate beliefs over the distribution of search results on the next page. We introduce a Bayesian updating approach to the model.

Previous search cost literature sets beliefs to reflect the actual distribution occurring in the data. We relax this assumption by introducing learning to occur over the distribution of prices for hotels in a given star-rating category. The distribution of hotel characteristics is still assumed to be drawn from the empirical distribution of hotels occurring on the second page. Instead of taking the prices from the empirical distribution, however, we assume agents learn about the average price of hotels belonging to each star-rating from the realized first period draws in a Bayesian manner. Details of implementing learning are contained in the section on estimation.

3.2.3.1 Discussion on Learning

The use of the empirical distribution to form beliefs is synonymous to an agent knowing all available offers prior to search, but not knowing which offers will be made to her, in which period they will be made, and what individual taste shock, ε_{ij} she will experience. For most search applications, especially those involving differentiated products, this assumption is likely unreasonable. Though we limit learning to the mean of prices by star-rating category to simplify computation, we believe this captures agent behavior quite well.

The price distribution constitutes a good starting point for learning as it is the most time variant from search session to search session, and would require learning even from seasoned travelers to the area. We conservatively restrict learning to occur within star-rating categories, in order to condition prices on the hotel characteristics. In practice, this lowers the variance of price draws matched to rooms in any category. Characteristics are taken from the empirical distribution. These assumption may be most appropriate for travelers who have previously visited the area, or similar areas. Future work may extend learning to include both the mean and variance of prices, or a full joint distribution of characteristics and prices.

Other processes of learning are likely to be important as well. Agents may be learning about the star-rating cutoff values along a continuum of quality for the area. It is likely that learning about the price distribution under a particular star category influences beliefs about the distribution for other categories as well. For now, we assume that prices are assigned based on quality and market factors, that the quality will be highly correlated with the star rating, but that the star rating does not directly influence the prices set.

3.3 Data

We use a unique data set that captures the search and action histories of agents visiting a popular hotel search website. The set of search request parameters, such as the dates for the reservation, the number of guests, whether the reservation is over a weekend, and the number of days searching in advance of the visit are observed. The data also include all search results observed by the agent and the order the results, which are hotel rooms and their prices, are presented on a page. Fifteen hotels are presented on each page viewed. Finally, we also observe the history of actions taken by the agent. The data contain whether or not the agents revisit pages, click on a hotel option, or engage in sorting or filtering of results. We observe all agents searching for hotels in the city of Chicago for the month of May, 2007.

In this work, we restrict attention to default sorting agents, those who do not do any sorting or filtering of results. We also take a click as an indication of demand, as clicking a hotel option takes the agent to the hotel's external website to book the reservation which we cannot observe. When an agent makes multiple clicks, we take the last click as their selected option. There does exist some data using individual agent IDs, in which the agent has registered a profile with the website. We do not include these data in the analysis at this point, though may consider it in future applications. Further discussion of the data, including descriptive statistics, may be found in Koulayev (2014).

3.4 Estimation

We will estimate the model via simulated maximum likelihood. The structural assumptions over the utility shock allows us to construct the likelihoods of the click decisions conditional on observed hotel characteristics that resemble familiar logit discrete choice model expressions. We introduce randomly drawn search costs and estimate the distribution parameters following a mixed logit model similar to those used to estimate models with random coefficients. We proceed by first stating the above mathematically, followed by a brief outline of how we construct the various likelihoods accounting for search and return, and conclude with a thorough discussion of each step.

In equations (6) and (7) we state the agent's utility function for a given hotel option, and break that down into the sum of mean utility (observed) and the utility shock which is unobserved by the econometrician. Assuming the utility shock is distributed i.i.d Extreme Value Type 1, the likelihood that agent i chooses a particular hotel and length of search may be expressed in the familiar logit model framework with the length of search and the page selection as arguments conditioned on search costs and mean utilities of the options. The exact likelihood expressions for each joint set of search length and page choice are provided in Appendix I. Since each agent experiences different choice sets, the likelihoods are more complicated than those in standard logit models.

The resulting likelihood is deterministic for a given set of taste coefficients, beliefs over the next period values, and a given search cost. If the exercise were to estimate the set of taste coefficients, from here we could sum the log of respective likelihoods of each history. However, we assume heterogeneous search costs, and wish to learn about the distribution of these costs. To do so, we will draw search costs from a log normal distribution and estimate the set of taste coefficients as well as the mean and variance of the search costs via a mixed logit specification. The likelihood for an agent with search length t and page selection k is given by,

$$P_i(k,t) = \int L_{k,t;i}(k,t|\theta,c)h(c|\mu,\sigma)dc \tag{3.8}$$

The integral must be approximated and is done so by simulation, within the likelihood maximization routine. In the above, θ, μ, σ are being estimated.

Outline of estimation strategy For each iteration of the optimization routine, complete the following steps:

1. Set initial parameter values for the vector θ, and for parameters entering belief updating. The mean utilities from the observed options on the first page may then be calculated.

2. Update beliefs over the distribution of next page hotels. Simulate draws from that distribution and calculate the mean utilities present on each simulated page. Using Lemma 1, derive the distribution of the maximal utility option on each page, take a draw from that distribution, and store that utility in an $S \times 1$ vector for each agent, denoted B_i.

3. Set initial parameter values for the mean and variance of a log-normal distribution. Draw a set of search costs from that distribution.

4. For each search cost draw, and using the simulated utilities from the next page, compute the set of reservation utilities.

5. Taking each search cost draw as given, compute the likelihoods using the mean utilities and reservation values.

6. Average the likelihoods across simulated search costs, then sum the logs of the averaged likelihoods.

The optimization occurs over the taste parameters θ, and distribution parameters μ, and σ^2 for the search costs, and the parameters used in belief updating.

3.4.1 Specifics

For the two period model, we consider the set of histories that involve either no search or search once and stop. Taking the set of parameters, θ, as given, we can compute the mean utility of each hotel option, μ_{i1}^h for $h = 1, ..., 15$ from the observable characteristics of that hotel, for each first page of these histories, indexed as agents $i = 1, ..., N$.

For each first page, an agent must decide whether or not to engage in search. The benefit from search is unknown to the agent at this point, but the expected benefit depends on their beliefs over the distribution of hotel characteristics and match values from which the results will be drawn. To approximate the expected payoff, we model beliefs and simulate possible second page results and match values and average over them.

3.4.1.1 Modeling beliefs

The most flexible specification for belief updating would allow for learning to occur over the joint distribution of all characteristics and prices with regards to means, variances, and covariances. Such a formulation would impose a significant computational burden, and may not drastically improve estimation above a much more restrictive view of learning. Therefore, we limit the current consideration for learning to occur over price means for a given star-rating. We choose the star-rating because it is an important determinant of price and is likely to be the most cognitively simple characteristic by which agents may categorize hotels. Other important factors such as the distance from the airport or city center are measured continuously and would pose a greater modeling challenge. Table 3.3 presents the results from a regression of prices on hotel characteristics. The signs on coefficients on all characteristics match intuition, and prices increase in star-ratings at a decreasing rate. The neighborhood indicators also register as important factors in price, and may present an alternative dimension by which to allow learning to occur.

We assume that agents learn about the mean of prices independently across star-ratings, and take the respective variances as given. We suppress the notation for the star-rating category in what follows to prevent clutter. However, the process should be understood to occur separately for each star-rating category. Learning occurs according to Bayes' rule. Agents believe that prices are normally distributed with an unknown mean θ_p and known variance σ_p^2. We set the variance to match that occurring in the actual distribution of prices for all observed first page results for each respective star-rating. The following steps regarding updating occur for each star-rating category independently. Notation and analysis are consistent with that in Poirier (1995). As in Ackerberg (2003) and Crawford and Shum (2005), we will exploit the set of natural conjugate priors to greatly reduce the computation for deriving posterior beliefs.

All agents enter the first page of results with the same prior beliefs of the unknown mean,

$$\theta_p|\theta_v \sim N(\underline{\mu}, \underline{h}^{-1}) \tag{3.9}$$

where $\theta_v \equiv \frac{1}{\sigma^2}$ is known as the *precision* and, $\underline{\mu}$ is set to match the mean observed over first page results, and $\underline{h} > 0$ determines the strength of agent belief in the prior mean. We will set this value to 2 for now but may estimate its optimal value in the future. We introduce notation for what follows. Let,

$$h = [\theta_v^{-1}/T]^{-1} = T\theta_v \tag{3.10}$$

which will determine how the precision gets updated. T is the number of hotel observations in the agent's first page draws that fall in the respective category.

$$\overline{h} = \underline{h} + h \tag{3.11}$$

expresses the updated precision, and

$$\overline{\mu} = \overline{h}^{-1}(\underline{h}\underline{\mu} + h\overline{y}) \tag{3.12}$$

expresses the updated mean. We also introduce the following notation that will appear in some of the calculations below,

$$v = T - 1 \tag{3.13}$$

$$s^2 = v^{-1}\sum_{t=1}^{T}(y_t - \overline{y})^2 \tag{3.14}$$

Bayes' Rule Recall Bayes' rule, as we apply it to a general density function, letting $f(\cdot)$ denote the prior density here ,

$f(\theta|y) = \frac{f(\theta,y)}{f(y)} = \frac{f(\theta)\mathcal{L}(\theta;y)}{f(y)}$

$$\propto f(\theta)\mathcal{L}(\theta; y) \tag{3.15}$$

in what follows, we will focus predominantly on the numerator because the denominator does not include the variable of interest. The denominator can easily be computed using,

$f(y) = \int \mathcal{L}(\theta; y)f(\theta)d\theta$

where y denotes observed draws from the distribution.

Deriving Update For each agent, we take the price draws from the first page as the observed data. The draws are considered to be iid, so the likelihood of θ_p given the observed data y and known precision θ_v is,

$\mathcal{L}(\theta_p; y|\theta_v) = \prod_{t=1}^{T}\phi(y_t|\theta_p, \theta_v^{-1})$

$= (2\pi\theta_v^{-1})^{-T/2}exp[-\frac{\theta_v}{2}\sum_{t=1}^{T}(y_t - \theta_p)^2]$

$= (2\pi\theta_v^{-1})^{-T/2}exp[-\frac{h}{2T}[vs^2 + T(\overline{y} - \theta_p)^2]]$

$$= b_1(\theta_v)\phi(\overline{y}|\theta_1, h^{-1}), \tag{3.16}$$

where

$b_1(\theta_v) = (2\pi)^{-1/2v}T^{-1/2}\theta_v^{v/2}exp(-\frac{1}{2}\theta_v vs^2)$ and does not depend on the unknown. Details on going from the second to third equality are in the appendix. We can then derive an expression for the numerator, (15),

$f(\theta_p|\theta_v)\mathcal{L}(\theta_p;y|\theta_v) = \phi(\theta_p|\underline{\mu},\underline{h}^{-1})b_1(\theta_v)\phi(\overline{y}|\theta_1,h^{-1})$

$= b_1(\theta_v)(2\pi\underline{h}^{-1})^{-1/2}(2\pi h^{-1})^{-1/2}exp[-\frac{1}{2}\{\underline{h}(\theta_p-\underline{\mu})^2 + h(\overline{y}-\theta_p)^2\}]$

$= b_1(\theta_v)(2\pi\underline{h}^{-1})^{-1/2}(2\pi h^{-1})^{-1/2}exp[-\frac{1}{2}\{\overline{h}(\theta_p-\overline{\mu})^2 + (\underline{h}^{-1}+h^{-1})(\overline{y}-\theta_p)^2\}]$

$$= b_1(\theta_v)\phi(\theta_p|\underline{\mu},\underline{h}^{-1}+h^{-1})\phi(\theta_p|\overline{\mu},\overline{h}^{-1}) \tag{3.17}$$

Using the above expression, we can derive the denominator by integrating out θ_p,

$f(y|\theta_v) = \int_{-inf}^{inf} f(\theta_p|\theta_v)\mathcal{L}(\theta_p;y|\theta_v)d\theta_p$

$= b_1(\theta_v)\phi(\overline{y}|\underline{\mu},\underline{h}^{-1}+h^{-1})\int_{-inf}^{inf}\phi(\theta_p|\overline{\mu},\overline{h}^{-1})d\theta_p$

$$= b_1(\theta_v)\phi(\overline{y}|\underline{\mu},\underline{h}^{-1}+h^{-1}) \tag{3.18}$$

We can now recover the posterior density of θ_p, taking θ_v as given, by dividing (17) by (18)

$$f(\theta_p|y,\theta_v) = \phi(\theta_p|\overline{\mu},\overline{h}^{-1}) \tag{3.19}$$

Thus, $\overline{\mu}$ from equation (12) is interpreted as the posterior mean, and $\overline{h}$ is the posterior precision. The value $\overline{\mu}$ will be used as the mean for simulating hotel draws on the second page. The assumption that no learning occurs in the variance will impose that the first page variance be used for second page draws. These two characteristics constitute departures from the literature which typically imposes the empirical distribution. Thus, beliefs for price draws for agent i on hotel j will be taken from,

$$p_{ji} \sim N(\overline{\mu}_i^{star_j}, \sigma_p^2) \tag{3.20}$$

where the subscript i reflects the fact that each agent will update to a new mean based on the hotel and price draws on their first page of results. The superscript on the mean denotes that a separate mean will exist for each star-rating.

3.4.1.2 Simulating second pages

To simulate the second page results for each first page, we construct S hypothetical second pages. For each hypothetical page, we draw $h_{s2}^1, h_{s2}^2, ..., h_{s2}^{15}$ from the empirical distribution of second page hotels. To ease the computational burden, we allow hotels to be drawn with replacement according to the frequency of their appearance on observed second pages. This speeds up computing time drastically and does little to compromise the actual estimation. An $i.i.d.$ shock is attached to each drawn hotel and the price paired to each hotel will be randomly drawn from a distribution. For each drawn hotel on page s, we draw a corresponding price from the distribution of prices for that hotel star-rating category as described above. Once we have the characteristics and prices for each option on the simulated pages, the mean utility of each option can be calculated. Using Lemma 1 proven in Appendix I, the maximum utility from a simulated page is itself a random variable from the Extreme Value Type I distribution, with location parameter $M^s = log[exp(\mu(p_1^s, q_1^s) + ... + \mu(p_{15}^s, q_{15}^s)]$, and unit scale.

Thus, once we have the mean utilities for the fifteen options on a page, the match value for the maximum utility (though it is unknown which of the fifteen would yield the maximum) may be drawn from the distribution described above. Then the maximum utility from a simulated page, s, is computed $u_{max}^s = M^s + \varepsilon^s$. These maximum utilities can then be collected into an individual specific $S \times 1$ vector, B_i.

3.4.1.3 Search costs

We draw a set of random search cost values to be used as simulated points for each agent. We assume search costs are drawn from a log-normal distribution, and set initial values for the distribution parameters, mean μ and variance σ^2. We draw MS search cost values for each agent i, c_i^l, for $l = 1, ..., MS$. This set of search costs are then paired to simulated second page maximal utilities for each first page.

Taking this set of search costs and the beliefs over the next period's maximal utility, we can compute a reservation utility for each agent-search cost pair. The reservation utilities must be individual specific because the beliefs over the future are at the agent level when incorporating learning, as are the expected benefits from search. As argued in section (3.1.2), for a given search cost c_i^l, there exists a unique level of utility from the first page options denoted r_{1i}, such that the consumer is indifferent between searching and not,

$Q_1(r_{1i}) - r_{1i} = c_i^l$

Taking the future expected benefit from search as fixed, the reservation utility can be viewed as a function of the search cost. Thus, search will only occur in the two-period model when $u_{1i}^* < r_{1i}(c_i^l)$. The subscript on the reservation utility function reflects the learning that occurs at the agent level, yielding an agent specific expected benefit from search, and thus an agent specific reservation utility function. If this inequality is violated, then no search occurs and the agent chooses an option from the first page or the outside option. Substituting in the expected value functions derived earlier, we have

$Q_1(u_{1i}^*) - u_{1i}^* = c_i^l$

$\rightarrow E_{1i} max(u_{2i}, u_{1i}^*) - u_{1i}^* = c_i^l$

$\Rightarrow u_{1i}^* = r_{1i}$

which is solved using an approximation for the expected value. Note that the subscript on the expectations operator, E_{1i} indicates that expectations are unique to each agent.

3.4.1.4 Optimizing the likelihood

From the mean utilities of observed hotel options and the computed reservation utilities, we can construct the likelihoods as described in the Appendix, based on the observed search decision and outcome. This procedure yields MS likelihoods for each of N agents, denoted $L_{k,t;i}(k,t|\theta,c)$ in equation (3.8). Next, we approximate the integral in equation (3.8) by averaging the likelihoods over the MS draws of search costs for each agent,

$$\check{P}_i(k,t) = \frac{1}{MS}\sum_{l=1}^{MS} w(c_i^l) L_{k,t;i}(k,t|\theta,c_i^l) \tag{3.21}$$

where the drawn search cost nodes and the assigned weights $w(c_i^l)$ are consistent with Legendre nodes in a Gaussian quadrature.

We then sum over the logs of these N averages to compute the likelihood for our set of parameters, θ, μ and σ^2. Finally, we will optimize the likelihood with respect to our parameters using numerical methods.

3.5 Results

3.5.0.5 Beliefs

We first construct the priors by star-rating to be shared by all agents. We assume agents believe prices follow a normal distribution within each star-rating category. The mean and variance of prices within each star-rating are presented in Table 3.1. The precision is also listed, since the updating is calculated in those terms. The mean increases with star rating as would be expected. A greater dispersion of prices is found for three and four star hotels, which also constitute the bulk of of the observed hotels. These distribution parameters are taken from all default sorters in the data, not just those restricting search to the first and second pages.

Table 3.2 presents the mean, variance and precision of prices occurring on the second page of results. In general, the hotels with three, four and five star-ratings appearing on the second page have slightly higher prices on average than their first page counterparts, and are slightly more tightly distributed. The prices of one and two star hotels decline in average slightly. This pattern suggests that the website may be selecting hotels from all categories that more closely resemble a median-type of quality and price hotel from the entire distribution of hotels. The second page is then populated by the outer tails of these within star-rating distributions.

The results of the Bayesian updating are also presented in Table 3.2. We present the averages and standard deviations for the computed mean and precision across agents, since beliefs are agent specific. The average beliefs on the mean of the price distribution overestimate the empirical mean for one to three star-rated hotels and underestimate the means for the four and five star hotels. This is consistent with the difference in mean prices between the actual distributions for first and second pages described above.

The difference between the precision presented for the empirical distribution and that computed for the beliefs falls within a reasonable range. The former is calculated as the inverse of the empirical variance. The latter is a weighted average of the prior precision and the precision calculated from the agent's observed hotels and prices. The level of confidence in the prior precision influences the weighting and must be set exogenously, and we have assigned what we consider to be a low level of confidence.

3.5.0.6 Pricing and characteristics

The results of a regression of all hotel prices appearing in the set of default sorters on hotel characteristics are presented in Table 3.3. All estimates are precisely estimated and exhibit the anticipated sign. The estimates suggest that price decreases at an increasing rate with distance to the city center. The large positive correlation between price and distance from O'Hare Airport is perhaps a bit surprising given that we also control for neighborhood and distance from the city center. However, neighborhoods are large, and additional distance from the airport likely reduces noise and other negative factors associated with being close to the airport. Further, the most popular hotels are known to be located close to the city center, which is far from the airport. The small, yet precisely estimated, coefficient on distance from the city center suggests some multicollinearity may exist between these two measures. Importantly for our strategy to segment learning, the star-rating of a hotel has a strong influence over prices. Prices increase at a decreasing rate with the star rating.

3.5.0.7 Search cost estimates

We estimate a restricted form of the model. Fixing the utility parameters for price, star-rating, and distances from the city center and O'Hare Airport, we estimate the mean and standard deviation parameters for the distribution of search costs. The coefficients on indicator variables for neighborhoods and hotel chains are set to zero. In this simplified model, we also use the same search cost nodes for every agent. The simulated hotels and prices for the second page are still unique for each agent.

The median search cost is estimated to be \$36.92.[1] This median is relatively higher than recent estimates of search costs (De los Santos et al., 2013; Koulayev, 2014), though is estimated under a fairly restrictive specification of the model. Limiting attention to first and second page searchers may also result in over-sampling agents from the upper end of the search cost distribution for the whole population, explaining the high estimate. Optimizing over the taste parameters, or relaxing the assumption that learning occurs strictly within star ratings, would place less weight on search costs to explain agent search behavior. The results for the estimated search cost parameters are presented in the column denoted as with-learning in Table 3.5, along with the fixed parameters used for taste coefficients. The likelihood function behaves nicely along both dimensions of estimated parameters. Figures 3.1 and 3.2 show the likelihood functions graphed around the estimate of the mean and variance for the model with learning, respectively, holding the other estimated value at its optimum.

[1]The median search cost is calculated by taking the exponent of the estimated mean from the distribution, dividing by the coefficient on price, and multiplying it by 100 to scale it back to dollars.

To provide a direct comparison of estimated search costs from the model with learning, we run the estimation procedure with beliefs based instead on the empirical distribution of second page prices. The median search cost from the model without learning is almost a dollar higher. While this difference is small in magnitude, but still statistically significant, much of the similarity may be attributed to the chosen specification. In the two period model, learning occurs only once. Further, the assumptions of a rational prior and learning within star ratings produces a conservative environment for the degree of learning made possible. Learning over multiple periods of search, looser attention to star ratings, and different selections of priors may all result in a divergence in the estimates from the model with and without learning. Finally, the estimated parameter for the standard deviation of search costs is slightly larger in the model with learning.

3.6 Counterfactual

Motivated by the interest of on-line retailers in the positioning of their products or advertisements on web pages and within search results, we consider the effects of product placement on demand. We explore the effect of hotel placement on demand in a counterfactual that populates all first page results with the top fifteen most clicked hotels by default sorters.

First, we create a set of simulated agents from the empirical distribution of default sorters. The number of simulated agents matches the actual number of default sorters in the data. This comprises of simulated first and second page results from their respective empirical distributions, and draws for the outside option from the set of default sorter requests and estimated coefficients for each request characteristic. The simulated first and second page hotel draws are done without replacement for each agent. The set of beliefs for each simulated agent are constructed using our model with belief updating, search costs are drawn randomly from the log normal distribution with the estimated parameters, and reservation utilities are calculated. We then assign Extreme Value Type I shocks to each hotel and compute overall utility with the estimated coefficients on hotel characteristics and price. Demand is determined by comparing the utilities of the outside option, the greatest first page utility, and if the reservation utility is not met, the greatest second page utility. If the greatest utility option comes from the first or second page, then demand for that agent is set to one.

From the set of simulated default sorters, we draw 2,012 agents to match the number of actual agents who view at most two pages of results. We repeat these draws 100 times, each time assigning new random shocks and randomly selected outside option utilities. For the 100 draws, we find an average demand of 86.64 with standard deviation 7.17.

We replicate this process but now allow every agent to face the same first page results which are populated by the fifteen hotels with the highest amount of clicks by default sorters. The price of each hotel is drawn from the hotel's empirical distribution of prices. For the 100 draws, we find an average demand of 95.50 with standard deviation 6.76. The histograms for these two counterfactuals are presented in Figure 3.3.

The mean of demand from the group facing the most popular hotels as their first page choices is statistically different and greater than that from the group facing the empirical distribution of first page hotels. The result suggests that the placement of alternatives within the set of search results carries economically significant welfare implications. The internet search site receives payment from the hotels for each click. If prices are uniform, this would imply revenues could be increased by placing the more popular offers on the first page. Second, the increase in demand suggests consumer welfare has increased as well.

Demand at 95.50 out of 2,012 agents constitutes a low-take up rate of 4.7%, with 86.64 being 4.3%. In the data we observe 644 agents out of 2,012 select a hotel from the first or second page, or 32% take up. This difference is likely driven by optimizing only over the parameters of the distribution of search costs. The coefficients for price, star-rating, and distances are taken from previous work, while the controls for neighborhood and chain affiliation are omitted entirely. A further abstraction in both counterfactuals is the separation of hotel offers and the search requests entered by the agent. Available hotels and prices are tied to the search parameters, but in the counterfactuals these search parameters are converted into the value of the outside option and assigned randomly. Thus, the counterfactual support is independent of the outside option, while it will not be in reality. Expanding the counterfactual design to condition the support of hotels on the search parameters could also help improve matching the take up rates.

A more subtle point is that most of the 2,012 actual agents had the opportunity to view a third page of results but chose instead to forgo search. In the framework of our model, their reservation utility had been satisfied. Thus, the actual 2,012 agents from the data are likely to have higher first and second page utility draws than the set of randomly drawn counterfactual agents. To overcome this selection issue, we could assume a third page of results is possible, construct beliefs at the 2nd page and compute reservations utilities. We could then draw our counterfactual sample from those agents meeting their reservation utility after two pages to more closely approximate the true data generating process.

To push these counterfactuals further, the first item would be to estimate the full set of parameters. That will likely bring the amount of clicked hotels to its proper level of 32%. Next, the policy of populating the first page with the top fifteen most popular hotels could be relaxed. Instead, the characteristics of the most popular hotels could be used to adjust the support from which first page hotels are drawn. This policy would allow new hotel offers fitting desired characteristics to be shown on first pages upon introduction. The most popular hotels are unlikely to have available rooms for every search, but constructing the support based on popular characteristics gives much more flexibility in populating the first page with highly desired rooms.

3.7 Conclusion

The traditional models of search assume agents draw alternatives from a distribution that is known to them, leading to properties that do not match observed behavior. In this chapter, we estimate a search model for differentiated goods, relaxing the assumption of a known distribution of alternatives. Rather than pursuing a non-parametric form of learning, such as Dirichlet priors, we assume agents learn about unknown parameters for the distribution of prices which are assumed to be distributed normally. The Bayesian updating of normal priors allows for a flexible form of learning, where agent-specific reservation utilities may increase or decrease based on the agent's search history. This flexibility is attractive as it allows reservation utilities to evolve in ways driven by the data rather than the assumed form of learning.

In estimating the search costs of agents visiting a popular online search website to shop for hotel rooms in Chicago in May of 2007, we find that the model with learning yields marginally lower search costs than one in which agents form beliefs from the empirical distribution. While the difference may not be large, estimating a specification in which learning occurs across multiple periods of search could result in potentially large effects from learning on search costs. We then use the model to evaluate a policy that would populate the first page results with the most popular hotel options. The number of hotels selected is shown to significantly increase under this counterfactual. This finding suggests that product placement in search results can have important implications for the market.

Relaxing the assumptions of the learning process would provide a deeper understanding into the effect of learning on search. The assumption of normality in the price distribution prior is strong; considering alternative distributions would provide insights to learning effects separate from the assumed parametric form. Different specifications of learning within and across star ratings, learning over joint distributions of prices and characteristics, and learning beginning from different values in the prior should also be explored. Finally, the model should be extended to accommodate multiple periods to evaluate the cumulative effects of learning over repeated search.

3.8 Appendix I

3.8.1 Identities for Belief Updating

The following identities are adapted from Poirier (1995).

Identity 1 $\sum_{t=1}^{T}(y_t - \theta_p)^2 = \sum_{t=1}^{T}(y_t^2 + \overline{y}^2 - \overline{y}^2 - 2y_t\theta_p + {\theta_p}^2)$

$$= \sum_{t=1}^{T}\{(y_t - \overline{y})^2 + 2y_t\overline{y} + (\overline{y} - \theta_p)^2 + 2\overline{y}\theta_p - y^2 - y^2 - 2y_t\theta_p\}$$

$$= \sum_{t=1}^{T}(y_t - \overline{y})^2 + T(\overline{y} - \theta_p)^2 + 2T\overline{y}\theta_p - T\overline{y}^2 - T\overline{y}^2 - 2\theta_p\sum_{t=1}^{T}y_t + 2\overline{y}\sum_{t=1}^{T}y_t$$

$$= \sum_{t=1}^{T}(y_t - \overline{y})^2 + T(\overline{y} - \theta_p)^2$$

$$= vs^2 + T(\overline{y} - \theta_p)^2$$

Identity 2 $\underline{h}(\theta_p - \overline{\mu})^2 + h(\overline{y} - \theta_p)^2 = \underline{h}[\theta_p^2 + \underline{\mu} - 2\theta_p\underline{\mu}] + h[\overline{y}^2 + \theta_p^2 - 2\overline{y}\theta_p]$

$$= (\underline{h} + h)\theta_p + \underline{h}\underline{\mu}^2 + h\overline{y} - 2\theta_p[\underline{h}\underline{\mu} + h\overline{y}]$$

$$= \overline{h}\theta_p^2 + \underline{h}\underline{\mu}^2 + h\overline{y}^2 - 2\theta_p\overline{h}\overline{\mu}$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 - \overline{h}\overline{\mu}^2 + \underline{h}\underline{\mu}^2 + h\overline{y}^2$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 - (\underline{h} + h)(\underline{h} + h)^{-2}(\underline{h}\underline{\mu} + h\overline{y})^2 + \underline{h}\underline{\mu}^2 + h\overline{y}^2$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 - (\underline{h} + h)^{-1}[(\underline{h}\underline{\mu} + h\overline{y})^2 - (\underline{h} + h)^{-1}\underline{h}\underline{\mu}^2 + (\underline{h} + h)^{-1}h\overline{y}^2]$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 - (\underline{h} + h)^{-1}[\underline{h}^2\underline{\mu}^2 + h^2\overline{y}^2 + 2\underline{h}h\underline{\mu}\overline{y} - \underline{h}^2\underline{\mu}^2 - \underline{h}h\underline{\mu}^2 - \underline{h}h\overline{y}^2 - h^2\overline{y}^2]$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 - (\underline{h} + h)^{-1}[2\underline{h}h\underline{\mu}\overline{y} - \underline{h}h\underline{\mu}^2 - \underline{h}h\overline{y}^2]$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 - (\underline{h} + h)^{-1}\underline{h}h(-1)(\overline{y} - \underline{\mu})^2$$

$$= \overline{h}(\theta_p - \overline{\mu})^2 + (\underline{h}^{-1} + h^{-1})^{-1}(\overline{y} - \underline{\mu})^2$$

Deriving Likelihood Functions

To perform the maximum likelihood estimation to recover the set of characteristic taste parameters θ and the parameters of the search cost distribution μ and σ, we must derive the likelihoods for joint search and clicking decisions. We derive likelihoods for which the unobserved utility shock is integrated out, following Koulayev (2014), greatly reducing the computational burden required to evaluate the model.

3.8.2 Extreme Value Distribution

Recall the form of the Extreme Value Type 1 distribution with location parameter a and a unit scale,

CDF: $F_x(x) = exp(-e^{-(x-a)})$

PDF: $f_x(x) = exp(-e^{-(x-a)})e^{-(x-a)}$

To center the CDF to a standard EV Type 1 distribution with location parameter zero and unit scale, denoted $F(\cdot)$, we simply subtract the location parameter from the argument, so that $F_x(x) = F(x-a)$.

Also, note

$$F(infinity) = exp(-e^{-(inf-a)}) = exp(0) = 1$$

and

$$F(-infinity) = exp(-e^{-(-inf-a)}) = exp(-infinity) = 0$$

3.8.3 Lemma 1

The distribution of the maximum draw of 15 independent EV Type I random variables with location parameters $\mu_1, ..., \mu_{15}$ and unit scale is itself EV Type I with location parameter $M(\mu_1, ..., \mu_{15}) = log(exp(\mu_1) + ... + exp(\mu_{15}))$ and unit scale.

3.8.3.1 Proof:

We may express the CDF of the maximum as $P(max(u_1, ..., u_{15}) < x) = F_{u_1}(x) \times ... \times F_{u_{15}}(x) = F(x-\mu_1) \times ... \times F(x-\mu_{15})$. This product of CDFs can be written as

$F(x-u_1) \times ... \times F(x-u_{15}) = exp(-e^{-(x-\mu_1)} - ... - e^{-(x-\mu_{15})})$

$= exp(-e^{-x}e^{\mu_1} - ... - e^{-x}e^{\mu_{15}})$

$= exp(-e^{-x}(e^{\mu_1} + ... + e^{\mu_{15}}))$

$= exp(-e^{-x}exp(log[e^{\mu_1} + ... + e^{\mu_{15}}]))$

$= exp(-e^{-(x-log[e^{\mu_1}+...+e^{\mu_{15}}])})$

$= exp(-e^{-(x-M(\mu_1,...,\mu_{15}))})$

$= F(x - M(\mu_1, ..., \mu_{15}))$

Lemma 2

Let x, y be independent EV Type 1 random variables with location parameters μ_x and a, respectively. Consider the pair of constants x_L and x_H, where $x_L < x_H$. Then the event: $x > y$, $x_L < x < x_H$ has probability:

$P(x > y, x_L < x < x_H) = \int_{x_L}^{x_H} F_y(x)f_x(x)dx$

$= \frac{exp(\mu_x)}{exp(M(a,\mu_x))}(F(x_H - M(a, \mu_x)) - F(x_L - M(a, \mu_x)))$

Proof:

We first center the CDF, then substitute in the definitions for the CDF and PDF and simplify,

$\int_{x_L}^{x_H} F_y(x)f_x(x)dx = \int_{x_L}^{x_H} F(x-a)f_x(x)dx$

$= \int_{x_L}^{x_H} exp(-e^{-(x-a)})exp(-e^{-(x-\mu_x)})e^{-(x-\mu_x)}dx$

$= \int_{x_L}^{x_H} exp(-e^{-(x-a)} - e^{-(x-\mu_x)})e^{-(x-\mu_x)}dx$

$= \int_{x_L}^{x_H} exp(-e^{-x}e^{a} - e^{-x}e^{\mu_x})e^{-x}e^{\mu_x}dx$

$= \int_{x_L}^{x_H} exp(-e^{-x}(e^{a} + e^{\mu_x}))e^{-x}e^{\mu_x}dx$

Now make the substitution, $t = e^{-x}$, $dt = -e^{-x}dx$

$= \int_{exp(-x_H)}^{exp(-x_L)} exp(-t(e^{a} + e^{\mu_x}))e^{\mu_x}dt$

$= -\frac{1}{e^{a}+e^{\mu_x}}exp(-t(e^{a} + e^{\mu_x}))e^{\mu_x} \,|_{exp(-x_H)}^{exp(-x_L)}$

$= -\frac{e^{\mu_x}}{e^{a}+e^{\mu_x}}[exp(-e^{-x_L}(e^{a} + e^{\mu_x})) - exp(-e^{-x_H}(e^{a} + e^{\mu_x}))]$

$= -\frac{e^{\mu_x}}{e^{a}+e^{\mu_x}}[F(x_L - a)F(x_L - \mu_x) - F(x_H - a)F(x_H - \mu_x)]$

$= \frac{e^{\mu_x}}{exp(M(a,\mu_x))}[F(x_H - M(a,\mu_x)) - F(x_L - M(a,\mu_x))]$

where the last equality holds by Lemma 1 and $M(a,\mu_x) = log(exp(a) + exp(\mu_x))$.

Property of $M(\cdot)$ From our definition of $M(\cdot)$, we have that $M(a,b) = log(exp(a) + exp(b))$. Then $M(M(a,b),c) = log\{exp(log[exp(a) + exp(b)]) + exp(c)\} = log(exp(a) + exp(b) + exp(c)) = M(a,b,c)$.

3.8.4 Inequalities and Relationships

To construct the likelihoods for joint search and clicking decisions, we must first build the relationships between the utilities of observed options. In what follows, we drop the consumer index for clarity in notation. Let,

t: number of pages observed by agent over the course of search, $t \in \{1,2\}$.

k: index for the clicked page. $k \in \{0,1,...,t\}$, where $k = 0$ denotes selecting the outside option.

u_0, u_1, u_2 : maximal utility on each page of results, where u_0 denotes the utility of the outside option.

x_k: utility of the clicked product. This implies that $x_k = u_k$.

y_k : utility of the second best utility on page k. General form, y.is the second best utility on a given page.

r_1: reservation utility to consider when deciding to search to see page 2.

$r_2 = -inf$ is assumed to ensure search ends after viewing the second page.

3.8.5 Click inequalities

We interpret "clicked" options following revealed preference methodology. We summarize the set of possible utility relationship inferences based on observed clicking behavior,

$$x_k \geq u_g \quad for\ g \in \{0,1\};\ k \in \{1,2\} \tag{3.22}$$

$$x_k > u_g \quad g = 1\ if\ k = 0;\ \ g = 2\ if\ k = 1 \tag{3.23}$$

$$x_k > y_k \quad for\ k \in \{1, 2\} \tag{3.24}$$

where the first inequality states that the utility from a selected option that is not the outside option must yield a utility greater than or equal to the maximal utility on the previously observed pages. Similarly, the second inequality states that the selected option must yield a higher utility than the maximal utility from pages observed after that page. Finally, the third inequality states that the utility of the selected option must yield a higher utility than the other options on that page.

3.8.5.1 Search inequalities

Considering only two periods greatly simplifies the type of search behavior that must be analyzed. When the optimal decision is to continue searching, logic implies $u_1^* < r_1$, while the decision to stop searching implies $u_1^* > r_1$. The relevant search decisions for the two-period model arise from three distinct cases and may be characterized by the length of search, t, and the index of the clicked page, k.

Case 1: $k = t = 2$

Consider first when the second best observed utility option is from the first page. We have stated already that if $t = 2$, we must have $u_1^* < r_1$. By the definition of u_1^*, this inequality also implies $u_1 < r_1$. Thus,

$$u_1 < r_1 \tag{3.25}$$

When the second best utility option is the outside option, we uncover a similar inequality,

$$u_0 < r_1 \tag{3.26}$$

and can treat the outside option as occurring on the first page going forward.

In the two period model, the optimal stopping rule once we've reached the second page is trivial and characterized by the inequality,

$$x_k > r_2 \tag{3.27}$$

Recall, we set $r_2 = -inf$ to ensure this inequality holds.

Case 2: $k \in \{0, 1\};\ t = 1$

In this case no search occurs, and we already stated the resulting simple inequality,

$$x_k > r_1 \tag{3.28}$$

Case 3: $k\{0, 1\};\ t = 2$

Since the best utility option is observed immediately the first two equations for Case 1 may be ignored. We are left with the inequality necessary for search,

$$x_k < r_1 \tag{3.29}$$

and the trivial decision to stop searching once we reach the terminal period, $t = 2$,

$$x_k > r_2 \tag{3.30}$$

3.8.6 Likelihoods

We derive the joint search and click likelihoods in two steps, using the above inequalities as building blocks. First we compute the conditional likelihoods where the joint search and click likelihood is conditional on the utility of the clicked product. We do this by integrating out the product specific error terms from the click and search inequalities. These likelihoods will be closed form functions of the utility of the selected option, the mean utilities of observed options, and the computed reservation utilities.

In the second step, we integrate out the selected option utility, x_k.

3.8.6.1 Conditional Likelihoods

Further Notation

Let $M_1 = log[exp(u_1^1) + ... + exp(u_1^{15})]$ denote the location parameter of the maximal utility on the first page, using the result from Lemma 1. Similarly, we can define $M_{g1:g2} = log(\sum_{g=g1}^{g2} exp(M_g))$ as the location parameter of the maximal utility on pages $g1, ..., g2$ combined. Note, $g1 \in \{0,1\}$ and $g2 \in \{1,2\}$ in the two period model. Finally, M_k^y denotes the location parameter of the maximal utility of non-clicked results on page k.

Computing the Conditional Likelihoods

Consider the set of events that must occur for a particular option to be chosen. Following the notation above, let us denote the utility of the chosen option as x_k. The utility of every other observed option must then be less than x_k, and x_k must surpass the reservation utility on the last page observed. The utility relationships have been stated and discussed in inequalities (28) through (36). From inequality (30), letting $x_k \equiv x$, we find,

$$P_{30}(x) = F(x - M_k^y)\ for\ k \in \{1,2\}$$

Then from inequality (29), we find,

when $k = 0$,

$$P_{29}(x) = F(x - M_1) \times F(x - M_2) = F(x - M_{1:2}),$$

and when $k = 1$,

$$P_{29}(x) = F(x - M_2)$$

When the selected option is from the second page, we must consider both the click and search inequalities (28) and (31). Given $x_k = x$, the probability of these two events holding is,

$$P_{28,31} = F(min(x, r_1) - M_1)$$

This probability accounts for the unselected options on the first page and decisions to search. The min function accounts for the fact that the maximal unselected option from the first page yields lower utility than the selected utility, and the maximal utility from the first page was lower than the reserve utility to induce search.

Similarly, the inequalities regarding the outside option for the case when search occurs and the selected option is on the second page, from inequalities (28) and (32), yield the following probability,

$$P_{28,32}(x) = F(min(x, r_1) - \mu_0)$$

Collecting the above probabilities of events, we get the conditional likelihood function, conditional on the utility of the selected option.

$L(x, k, t) = F(min(x, r_1) - M_1) \quad if\ k = 2$

$\times F(min(x, r_1) - \mu_0) \quad if\ k = 2$

$\times F(x - M_k^y) \quad if\ k \in \{1, 2\}$

Writing these collected options out in the cases to be considered we have,

Case 1: $k = t = 2$

$L(x, k, t) = F(min(x, r_1) - M_1)$

$\times F(min(x, r_1) - \mu_0)$

$\times F(x - M_2^y)$

Case 2: $k \in \{0, 1\};\ t = 1$

$L(x, k, t) = F(x - M_1^y)$

where $M_1^y = M_1$ if the outside option is chosen.

Case 3: $k \in \{0, 1\};\ t = 2$

$L(x, k, t) = F(x - M_2)$

$\times F(x - M_1^y)$

where $M_1^y = M_1$ if the outside option is chosen, and M_1^y includes μ_0 if $k = 1$.

Unconditional Likelihoods

Now we integrate out the utility of the selected option to derive the unconditional likelihoods as functions of the index for the selected page and search length. The unconditional likelihood may be expressed as,

$$L(k, t) = \int_{LB}^{UB} L(x, k, t) exp(-e^{-(x-\mu_x)}) e^{-(x-\mu_x)} dx \tag{3.31}$$

where the conditional likelihood $L(x, k, t)$ will vary according to page selection and search length. Also, the lower bound (LB) and upper bound (UB) will vary case by case. Thus, we will derive three different likelihood equations, one for each page selection and search length case. We will rely heavily on Lemma 2 for this section.

Case 1: $k = t = 2$

The unconditional likelihood in this case is,

$$L(x, k, t) = F(min(x, r_1) - M_1) \times F(min(x, r_1) - \mu_0) \times F(x - M_2^y)$$

When $min(x, r_1) = x$, we can use Lemma 1 to rewrite this as, $L(x, k, t) = F(x - M(\mu_0, M_1, M_2^y))$.

When $min(x, r_1) = r_1$, then $L(x, k, t) = F(r_1 - M(\mu_0, M_1)) \times F(x - M_2^y)$.

Then using Lemma 2, we can express the integral as,

When $x < r_1$, the bounds of integration will be from -infinity to r_1, then

$$L(k, t)|(x < r_1) = \frac{exp(\mu_x)}{exp(M(\mu_x, \mu_0, M_1, M_2^y))}[F(r_1 - M(\mu_x, \mu_0, M_1, M_2^y)) - F(-infinity)]$$

which, using the property of $M(\cdot)$, simplifies to,

$$L(k, t)|(x < r_1) = \frac{exp(\mu_x)}{exp(M_{0:2})} F(r_t - M_{0:2}) \tag{3.32}$$

and when $x > r_1$, the bounds of integration will be from r_1 to infinity, then

$$L(k, t)|(x > r_1) = F(r_1 - M(\mu_0, M_1)) \frac{exp(\mu_x)}{exp(M(\mu_x, M_2^y))}[F(infinity) - F(r_1 - M(\mu_x, M_2^y))]$$

similarly, using the property of $M(\cdot)$, simplifies to,

$$L(k, t)|(x > r_1) = F(r_1 - M_{0:1}) \frac{exp(\mu_x)}{exp(M_2)}[1 - F(r_1 - M_2)] \tag{3.33}$$

Combining the two equations, we get the integration over the full range of x,

$$L(k,t) = \frac{exp(\mu_x)}{exp(M_{0:2})} F(r_t - M_{0:2}) + \frac{exp(\mu_x)}{exp(M_2)} F(r_1 - M_{0:1})[1 - F(r_1 - M_2)] \tag{3.34}$$

Case 2: $k \in \{0,1\};\ t = 1$

The unconditional likelihood in this case is,

$$L(x,k,t) = F(x - M_1^y)$$

where $M_1^y = M_1$ if the outside option is chosen. We can infer by the fact that they did not engage in search that $x > r_1$, thus the bounds of integration will be from r_1 to infinity. Using Lemma 2,

$$L(k,t) = \frac{exp(\mu_x)}{exp(M_{0:1})}[F(infinity) - F(r_1 - M_{0:1})]$$

$$= \frac{exp(\mu_x)}{exp(M_{0:1})}[1 - F(r_1 - M_{0:1})] \tag{3.35}$$

using the same property of $M(\cdot)$ to simplify the expression of the location parameter. Note, in the case of selecting the outside option, $M_0^y = M_1$ as M_1 provides the location parameter of the second best option from the options available with no search. Similarly, $M(\mu_0, M_1^y)$ is the location parameter for the distribution of the second best utility option when the best option is from the first page given no search.

Case 3: $k \in \{0,1\};\ t = 2$

The unconditional likelihood in this case is,

$$L(x,k,t) = F(x - M_2) \times F(x - M_1^y)$$

where $M_1^y = M_1$ if the outside option is chosen, and M_1^y includes μ_0 if $k = 1$. We can infer that $x < r_1$ since the best utility option is on the first page of results, including outside option, but that that option still induced search when it was seen originally. Thus, the bounds of integration will be from - infinity to r_1. Using Lemma 2,

$$L(k,t) = \frac{exp(\mu_x)}{exp(M_{0:2})}[F(r_1 - M_{0:2}) - F(-infinity)]$$

$$= \frac{exp(\mu_x)}{exp(M_{0:2})}F(r_1 - M_{0:2}) \tag{3.36}$$

3.9 Appendix II: Tables and Figures

Table 3.1: Priors on price distribution by star-rating, shared by all agents

Prior	Star Rating				
	1	2	3	4	5
Mean	0.8783	1.6532	2.4157	2.9328	5.022
Variance	0.1004	0.5634	1.5388	1.6985	2.6902
Precision	9.9559	1.7751	0.6498	0.5887	0.3717
Observation	1,375	70,957	135,262	126,656	1,932

Note: All default sorters used to calculate priors and posteriors. Total observations = 23,959.

Table 3.2: Second page price distribution parameters by star-rating, posterior values averaged over individual agent beliefs

	Star Rating				
	1	2	3	4	5
Empirical Paramters					
Mean	0.8158	1.1108	2.2564	3.021	5.2638
Variance	0.0651	0.3722	1.4319	1.4352	1.6653
Precision	15.356	2.6868	0.6984	0.6967	0.6005
Observations	7,914	47,577	58,141	49,570	1,652
Posterior Mean					
Mean	0.8804	1.7211	2.404	2.9278	5.0221
Standard deviation	0.0606	0.359	0.5046	0.5147	0.0685
Posterior Precision					
Mean	2.5714	7.257	5.6687	5.1123	2.03
Standard deviation	3.5078	4.3965	1.6646	1.6813	0.1184

Note: All default sorters used to calculate priors and posteriors. Total observations = 23,959.

Table 3.3: Price regression estimates from all default sorters

	Dependent variable: Price/100	
	Coefficient	SE
Distance (city-center)	-0.192	0.0015
Distance (city-center) squared	0.005	0.0001
Distance (O'Hare)	1.538	0.0132
Star-rating	0.561	0.0089
Star-rating squared	-0.023	0.0015
Neighborhood 2	-1.277	0.0155
Neighborhood 3	-2.664	0.0267
Neighborhood 4	-4.411	0.0394
Chain hotel 2	-0.440	0.0053
Chain hotel 3	-0.165	0.0035
Chain hotel 4	-0.152	0.0089
Chain hotel 5	-0.419	0.0116
Chain hotel 6	-0.703	0.0109
N = 721,848		

All prices faced by default sorters considered, for all pages. Non-robust standard errors.

Table 3.4: Search and selection in the sample used for estimation

Total Agents	2012
Agents view	
1 page	717
2 pages	1295
Agents select from	
1st page (non-outside option)	283
2nd page	361
Outside option	1368
Agents clicked a hotel	644

Table 3.5: Search Cost Estimates

Fixed Parameters		
Price	-0.73	
Distance	-0.55	
Distance O'Hare	-0.24	
Star Rating	0.32	
Search Cost Estimates		
	Baseline	With Learning
Log Normal Mean	-1.2879	-1.311
	[0.0014]	[0.0015]
Log Normal Std dev	1.716	1.728
	[0.0006]	[0.0006]
Median Search Cost	37.79	36.92
Log-likelihood	6,311	6,301

Maximum likelihood estimates are reported for the parameters of the log normal distribution. Standard errors are reported in brackets. The taste coefficients are fixed.

Figure 3.1: Likelihood function around estimated mean

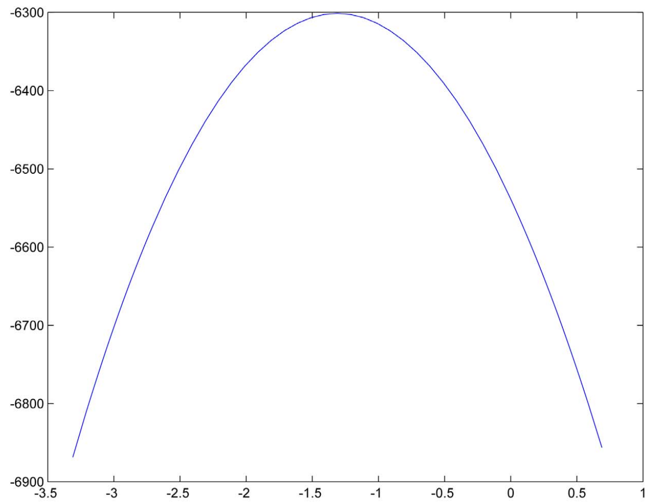


Figure 3.2: Likelihood function around estimated standard deviation

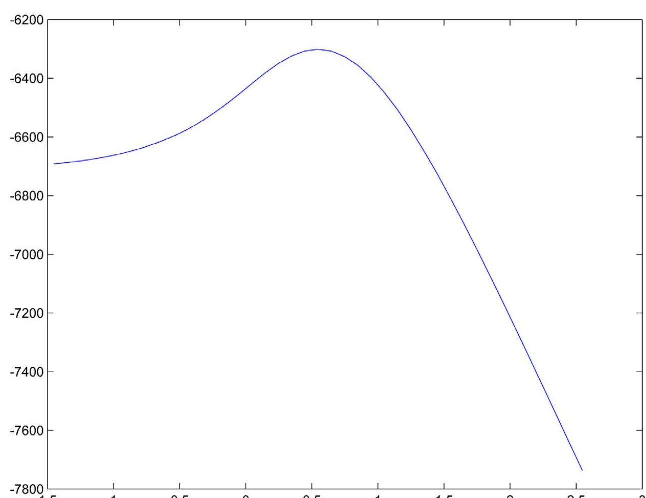
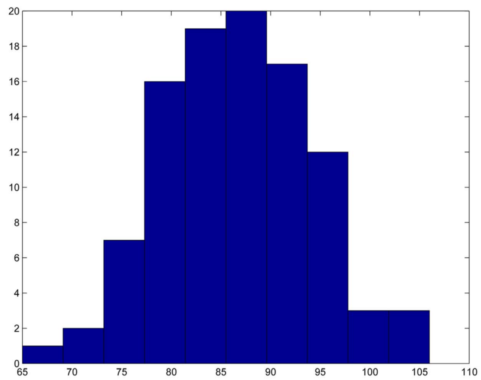
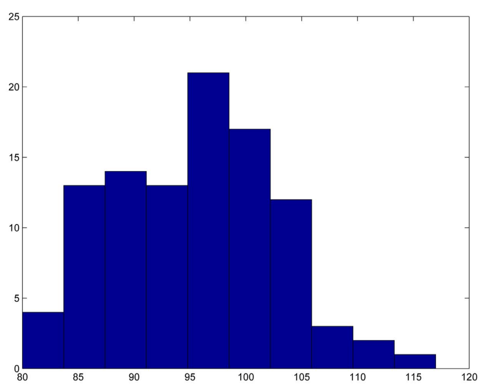


Figure 3.3: Histograms of counterfactual demand

(a)



(b)



Note: Panel (a) shows the histogram for demand by agents facing simulated first pages with draws from the empirical distribution of first pages for default sorters. Panel (b) shows the histogram for demand by agents facing identical simulated first pages populated by the 15 most popular hotels for default sorters. Each simulation includes 2,012 agents and 100 simulations are performed.

Bibliography

Ackerberg, D. A. (2003). Advertising, learning, and consumer choice in experience good markets: an empirical examination. *International Economic Review*, 44(3):1007–1040.

Aguirregabiria, V., Clark, R., and Wang, H. (2014). Diversification of geographic risk in retail bank networks: Evidence from bank expansion after the riegle-neal act. Technical report, CEPR Discussion Papers.

Angrist, J. D. and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575.

Banerjee, A. and Duflo, E. (2001). The nature of credit constraints: Evidence from an indian bank. In *Manuscript, MIT. Available at: http://www. chicagobooth. edu/research/workshops/AppliedEcon/archive/WebAr chive20012002/duflo. pdf*.

Banerjee, A. V., Cole, S., and Duflo, E. (2004). Banking reform in india. *India Policy Forum*, 1(1):277–332.

Banerjee, A. V. and Duflo, E. (2014). Do firms want to borrow more? testing credit constraints using a directed lending program. *The Review of Economic Studies*, 81(2):572–607.

Black, S. E. (1999). Do better schools matter? parental valuation of elementary education. *Quarterly journal of economics*, pages 577–599.

Bollard, A., Klenow, P. J., and Sharma, G. (2013). Indias mysterious manufacturing miracle. *Review of Economic Dynamics*, 16(1):59–85.

Burgess, R. and Pande, R. (2005). Do rural banks matter? evidence from the indian social banking experiment. *American Economic Review*, 95(3):780–795.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Chen, Y. and Yao, S. (2014). Sequential search with refinement: Model and application with click-stream data. *Available at SSRN 2130646*.

Cole, S. (2009). Financial development, bank ownership, and growth: Or, does quantity imply quality? *The Review of Economics and Statistics*, 91(1):33–51.

Cole, S., Giné, X., Tobacman, J., Topalova, P., Townsend, R., and Vickery, J. (2013a). Barriers to household risk management: Evidence from india. *American Economic Journal: Applied Economics*, 5(1):104–135.

Cole, S. A., Giné, X., and Vickery, J. I. (2013b). How does risk management influence production decisions? evidence from a field experiment. *Evidence from a Field Experiment (July 1, 2013). World Bank Policy Research Working Paper*, (6546).

Crawford, G. S. and Shum, M. (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4):1137–1173.

De los Santos, B., Hortaçsu, A., and Wildenbeest, M. R. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review*, 102(6):2955–2980.

De los Santos, B., Hortacsu, A., and Wildenbeest, M. R. (2013). Search with learning. *Available at SSRN 2163369*.

De Quidt, J., Fetzer, T. R., and Ghatak, M. (2013). Market structure and borrower welfare in microfinance. *Available at SSRN 2540048*.

Dehejia, R. and Lleras-Muney, A. (2007). Financial development and pathways of growth: State branching and deposit insurance laws in the united states, 1900–1940. *Journal of Law and Economics*, 50(2):239–272.

Fetzer, T. (2014). Can workfare programs moderate violence? evidence from india. Technical report, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Gelman, A. and Imbens, G. (2014). Why high-order polynomials should not be used in regression discontinuity designs. Technical report, National Bureau of Economic Research.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028.

Hong, H. and Shum, M. (2006). Using price distributions to estimate search costs. *The RAND Journal of Economics*, 37(2):257–275.

Hortaçsu, A. and Syverson, C. (2004). Product differentiation, search costs, and competition in the mutual fund industry: A case study of s&p 500 index funds. *The Quarterly Journal of Economics*, 119(2):403–456.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics*, 124(4):1403–1448.

Imbens, G. and Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, page rdr043.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Jayachandran, S. (2006). Selling labor low: Wage responses to productivity shocks in developing countries. *Journal of Political Economy*, 114(3):538–575.

Jayaratne, J. and Strahan, P. E. (1996). The finance-growth nexus: Evidence from bank branch deregulation. *The Quarterly Journal of Economics*, pages 639–670.

King, R. G. and Levine, R. (1993). Finance and growth: Schumpeter might be right. *The Quarterly Journal of Economics*, pages 717–737.

Klonner, S. and Oldiges, C. (2014). Safety net for india's poor or waste of public funds? poverty and welfare in the wake of the world's largest job guarantee program. Technical report, University of Heidelberg, Department of Economics.

Kochar, A. (2011). The distributive consequences of social banking: a microempirical analysis of the indian experience. *Economic Development and Cultural Change*, 59(2):251–280.

Koulayev, S. (2013). Search with dirichlet priors: estimation and implications for consumer demand. *Journal of Business & Economic Statistics*, 31(2):226–239.

Koulayev, S. (2014). Search for differentiated products: identification and estimation. *The RAND Journal of Economics*, 45(3):553–575.

Krishnan, K., Nandy, D., and Puri, M. (2014). Does financing spur small business productivity? evidence from a natural experiment. *Review of Financial Studies*, page hhu087.

La Porta, R., Lopez-de Silanes, F., and Shleifer, A. (2002). Government ownership of banks. *The Journal of Finance*, 57(1):265–301.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Lee, D. S., Moretti, E., and Butler, M. J. (2004). Do voters affect or elect policies? evidence from the u. s. house. *The Quarterly Journal of Economics*, 119(3):807–859.

McCall, J. J. (1970). Economics of information and job search. *The Quarterly Journal of Economics*, pages 113–126.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Mobarak, A. M. and Rosenzweig, M. (2014). Risk, insurance and wages in general equilibrium. Technical report, National Bureau of Economic Research.

Mortensen, D. T. (1970). Job search, the duration of unemployment, and the phillips curve. *The American Economic Review*, pages 847–862.

Nyarko, Y. and Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3):971–1005.

Panagariya, A. (2006). Bank branch expansion and poverty reduction: A comment. Technical report, Mimeo, Columbia University.

Poirier, D. J. (1995). *Intermediate statistics and econometrics: a comparative approach*. Mit Press.

Rajan, R. G. and Zingales, L. (1998). Financial dependence and growth. *The American Economic Review*, 88(3):559–586.

RBI (2009). Report of the group to review branch authorisation policy.

Rosenzweig, M. and Udry, C. R. (2013). Forecasting profitability. Technical report, National Bureau of Economic Research.

Rosenzweig, M. R. and Binswanger, H. P. (1993). Wealth, weather risk and the composition and profitability of agricultural investments. *The Economic Journal*, pages 56–78.

Rosenzweig, M. R. and Wolpin, K. I. (1993). Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: Investments in bullocks in india. *Journal of Political Economy*, pages 223–244.

Rothschild, M. (1974). Searching for the lowest price when the distribution of prices is unknown. *The Journal of Political Economy*, pages 689–711.

Salmon, T. C. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, 69(6):1597–1628.

Stigler, G. J. (1961). The economics of information. *The Journal of Political Economy*, pages 213–225.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309.

Townsend, R. M. (1994). Risk and insurance in village india. *Econometrica: Journal of the Econometric Society*, pages 539–591.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review*, 43(4):1249–1287.

Zimmermann, L. (2012). Labor market impacts of a large-scale public works program: evidence from the indian employment guarantee scheme.

Curriculum Vitae

Contact

Nathaniel Young

Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, USA

Education

Western Washington University, B.A., Economics and English Literature (Magna Cum Laude), 2006.

University of California at Davis, M.A., Economics, 2007.

Boston University PhD candidate, Present. Thesis advisor: Dilip Mookherjee.

Fellowships and Awards

Rosenstein-Rodan Prize, Institute for Economic Development, 2014

Research Grant, Center for Finance, Law and Policy, 2014

Summer Research Grant, Department of Economics, Boston University, 2011

Presidential Scholar from the College of Business and Economics, Western Washington University, 2006

Outstanding Graduate from the Economics Department, Western Washington University, 2006

Best Senior Project Proposal, with Research Grant, Western Washington University Honors Program, 2005

Working Papers

1. *Formal Banking and Economics Growth: Evidence from a Regression Discontinuity Analysis in India*, March 2015.
2. *Commercial Banking in Agricultural Areas: Smoothing Spatially Distributed Productivity Shocks Across Branch Networks*, March 2015.
3. *Estimating a Sequential Search Model with Bayesian Learning: A Case of Online Search for Differentiated Goods* (with Sergei Koulayev), March 2015.

Works in Progress

1. *Measuring Slum Evolution with Remote Sensing* (with Michael Gechter).