

Bernecker, Andreas; Boyer, Pierre C.; Gathmann, Christina

Working Paper

Trial and Error? Reelection Concerns and Policy Experimentation during the U.S. Welfare Reform

IZA Discussion Papers, No. 9113

Provided in Cooperation with:

IZA – Institute of Labor Economics

Suggested Citation: Bernecker, Andreas; Boyer, Pierre C.; Gathmann, Christina (2015) : Trial and Error? Reelection Concerns and Policy Experimentation during the U.S. Welfare Reform, IZA Discussion Papers, No. 9113, Institute for the Study of Labor (IZA), Bonn

This Version is available at:

<https://hdl.handle.net/10419/113986>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

IZA DP No. 9113

Trial and Error? Reelection Concerns and Policy Experimentation during the U.S. Welfare Reform

Andreas Bernecker
Pierre C. Boyer
Christina Gathmann

June 2015

Trial and Error? Reelection Concerns and Policy Experimentation during the U.S. Welfare Reform

Andreas Bernecker

University of Mannheim

Pierre C. Boyer

University of Mannheim and CESifo

Christina Gathmann

University of Heidelberg, IZA and CESifo

Discussion Paper No. 9113

June 2015

IZA

P.O. Box 7240

53072 Bonn

Germany

Phone: +49-228-3894-0

Fax: +49-228-3894-180

E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ABSTRACT

Trial and Error? Reelection Concerns and Policy Experimentation during the U.S. Welfare Reform^{*}

We study the political economy of policy innovations during the U.S. welfare reform in 1996. Specifically, we investigate how reputation concerns among governors influence the decision to experiment with welfare policies. In line with a political agency model, our empirical results suggest that governors with high reputation among the electorate are less likely to experiment with welfare policies than governors with low reputation. Yet, governors with high reputation who are less concerned about reelection actually experiment more than governors striving for reelection. Overall, our findings imply that reelection concerns may inhibit innovation in the public sector.

JEL Classification: I38, H11, H77, D78, D83

Keywords: policy innovation, reputation concerns, U.S. welfare reform, experimentation, reelection concerns

Corresponding author:

Christina Gathmann
Department of Economics
Alfred-Weber Institut
University of Heidelberg
Bergheimerstraße 20
69115 Heidelberg
Germany
E-mail: christina.gathmann@awi.uniheidelberg.de

^{*} We thank Toke Aidt, Manuel Arellano, Henning Bohn, Matz Dahlberg, Peter Fredriksson, Vincenzo Galasso, Björn Tyrefors Hinnerich, Mario Jametti, Helene Lundqvist, Andreas Madestam, Dylan Minor, Espen Moen, Jörg Oechssler, Paanu Poutvara, Torsten Persson, Per Pettersson-Lidbom, Federico Revelli, Christian Riis, Elisabeth Schulte-Runne, Albert Solé-Ollé, Christoph Vanberg, Galina Zudenkova and seminar participants at Freiburg, Heidelberg, IEB in Barcelona, Konstanz, Mannheim, Norwegian Business School, Stockholm, the Silvaplana Workshop on Political Economy, the Conference on Evaluation of Political Reforms in Mannheim, the European Public Choice Society Meeting, the CESifo Conference on Public Sector Economics, the IIPF, and the EEA Annual Meeting for helpful suggestions. We are very grateful to Greg Shaw, Dave Andersen, and Paul Ehmann for sharing some of their data, and the administrators of the Welfare Rules Database (WRD) at the Urban Institute for answering our questions. Jan Scherer, Michael Hellwig, and Robert Aue provided outstanding research assistance. We are responsible for all remaining errors.

1 Introduction

When President Clinton signed the ‘Personal Responsibility and Work Opportunity Reconciliation Act’ (PRWORA) in August 1996, it marked the most important reform of U.S. welfare policy since the New Deal. PRWORA abolished the entitlement to cash assistance for poor families, the ‘Aid to Families with Dependent Children’ (AFDC). The new ‘Temporary Aid to Needy Families’ (TANF) program defined welfare benefits as temporary assistance rather than as an entitlement to eligible families. At the same time, PRWORA decentralized the authority to design welfare policies to the individual states. After 1996, policy-makers in the states were largely free to choose eligibility criteria for benefit receipt, the requirements recipients had to fulfill while on welfare, or the type of sanctions imposed if recipients did not comply with these requirements (see, for example, Grogger and Karoly (2005), Moffitt (2008)).

A vast literature has analyzed the consequences of the 1996 welfare reform for caseloads, employment, earnings, poverty, marriage patterns and other socio-economic outcomes (see, for example, Bitler and Hoynes (2010) and Meyer and Sullivan (2008) for some recent evidence; and Blank (2002), Grogger and Karoly (2005), Moffitt (2001, 2002) for surveys). In contrast, the political economy of welfare reform in the U.S. states has hardly been investigated until now.¹ Many welfare rules had never been tried before and most states had little prior experience whether and how certain policies would turn out in their local jurisdiction. If little is known about the effectiveness of certain policies, there are potentially large gains from experimentation and possibly large (political) costs of failure.²

Yet, how much did states take advantage of the opportunity to experiment and learn about welfare rules before and after PRWORA? And more importantly, which factors helped to promote experimentation in the U.S. states during the welfare reform; and which inhibited innovation? This article provides a first answer to these questions. Such an analysis is important to better understand if and when innovations in the public sector occur, and why they seem to happen less frequently than desirable.

We argue that reelection concerns have had an important influence on the decision to experiment with new welfare policies and to possibly reverse an experiment later on. To illustrate these reputation concerns, we analyze a simple political agency model of policy experimentation, inspired by Majumdar and Mukand (2004). In the model, voters are uncertain about the best welfare policy but politicians may have some information depending on their prior political experience. Politicians care about voters’ welfare as well as their chances of reelection. During her term in office an incumbent can decide whether to stick to the known status quo policy or experiment with a new, uncertain

¹In contrast, several detailed studies trace the political process leading up to reform at the federal level in 1996; see Reintsma (2007), Weaver (2000), or Haskins (2006).

²The idea behind laboratory federalism is that states, in a world of imperfect information, engage in experimentation to learn about the best policy (see e.g. Oates (1999) for a survey).

policy. The experiment is ex-ante welfare-enhancing but may yield its benefits in the short or long run. After implementing the experiment, the incumbent observes its short-term outcome. If the experiment has no short-term benefit, it is either because the experiment was implemented in the wrong state of the world or because of bad luck. The politician can now decide whether to stick with the policy or revert back to the status quo. Prior to the next election, the final outcome of the policy choices is observed by both politician and voters.

A politician, uncertain whether the experiment is beneficial, weighs the (risky) reputation gains from an experiment against her safe initial reputation. The higher her initial reputation (and hence here reelection chances ex-ante), the more an incumbent has to lose if the policy experiment fails. Hence, reputation concerns make the politician hesitant to engage in risky experimentation and result in too little policy experimentation relative to the first-best. This reputation effect should be weaker for politicians who worry less about their reelection chances.

Our model also has something to say when a policy experiment might be reversed. If an experiment fails in the short term, it is either because the policy is wrong for the current state of the world or because of bad luck. A politician with little knowledge whether the experiment is beneficial in the current state of the world is more likely to make a mistake when experimenting than a politician who knows the state of the world. Yet, a policy reversal becomes less likely the larger the potential welfare gains of an experiment, which increases the incumbent's willingness to gamble with voters' welfare.

To test our predictions in the context of the U.S. welfare reform, we collect a new data set on welfare policy rules both before and after PRWORA. Our analysis thus spans both the period of welfare waivers states adopted under the old AFDC program as well as the policy changes implemented after the federal reform in 1996. We complement our panel of policy rules with detailed information on the political structure and socio-economic conditions in each state.

Our empirical results provide strong support for the idea that reelection concerns influence policy experimentation and reversals. A high initial reputation of the politician reduces policy experimentation both during the waiver and TANF period, especially among governors who may be reelected. The influence of reelection concerns on experimentation is economically sizable: An increase in initial reputation by an interquartile range reduces experimentation among governors that can be reelected by 40-45%. Furthermore, governors with relatively little experience are more likely to experiment when they are less concerned about reelection. Again, the effect is sizable: Lame ducks with relatively little experience are 43% more likely to experiment. We also find that governors with low experience are more likely to revert an experiment, but less likely when the potential gains of experimentation appear sufficiently high.

A number of robustness checks bolster our findings that reelection concerns matter for policy experimentation. Most importantly, one might expect that ideology influences the

decision to experiment as Republicans have long pushed for the idea of workfare instead of welfare. We find that neither the ideology of the governor nor the ideology of voters have an influence on experimentation. Our results are also robust to controlling for the influence of the legislature as measured by party composition, legislative polarization or electoral competition in a state. Politicians who have to decide whether to implement a policy might look to their neighbors for inspiration or engage in welfare competition instead. Defining various measures of “neighbors” in terms of geography, population size and ideology, we find some evidence that cross-state spillovers matter for policy experimentation. Finally, the demand for welfare might also create incentives or pressures to experiment. A state’s socio-demographic structure indeed matters, but has little effect on our main results. In sum, experimentation with waivers during the AFDC program and welfare rules during the TANF program both support the idea that reputation concerns are important to understand policy innovations during the U.S. welfare reform.

The paper is structured as follows. The next section discusses the related literature while Section 3 provides background information about the U.S. welfare reform. In Section 4, we introduce our political agency model and derive empirical predictions. We describe the data sources and empirical strategy in Section 5. Section 6 presents the main results and reports robustness checks. The last section contains concluding remarks.

2 Related Literature

Our analysis contributes to a growing empirical literature on the role of electoral concerns for policy outcomes. Most papers study the link between gubernatorial term limits and tax or spending decisions (see e.g. Besley and Case (1995a, 2003); List and Sturm (2006)). Gagliarducci and Nannicini (2013) show that higher earnings for mayors reduce local public spending largely because better pay attracts higher-quality politicians to run as mayors; the incentive component seems to play only a minor role in Italy. For the U.S., Alt et al. (2011) find however evidence that term limits exhibit both an accountability effect (where governors who can be reelected perform better than term-limited governors) and a competence effect (experience in office improves performance) where performance is measured by high economic growth and low spending, taxes and borrowing costs. A recent study by Ferraz and Finan (2011) demonstrates that electoral accountability reduces corruption of local politicians in Brazil. We focus instead on policy experimentation and learning from experimentation as our central outcome of interest. In our case, reputation effects may result in too little experimentation if the politician puts a high weight on reelection, and both the politician and the voter are uncertain about the benefits of the experiment.

Our analysis is related to a sizable political science literature on policy innovation and diffusion which has studied policies such as the adoption of state lotteries or labor market regulation (e.g. Walker (1969), Cnudde and McCrone (1969), Allen and Clark

(1981), Berry and Berry (1990, 1992), or Boehmke and Skinner (2012); see Karch (2007) and Berry and Berry (2007) for surveys). Instead, we investigate the political economy of the U.S. welfare reform which is considered one of the most important social reforms since the New Deal. Furthermore, we stress reputational concerns among governors as an important mechanism, both theoretically and empirically, for policy innovations.³

Our theoretical setup is related to a large literature on reputation concerns following the seminal work of Holmström (1982, 1999). An early application to the political arena is the political agency model by Rogoff (1990). Closer to our setting are models where voters and politicians are both uncertain about the mapping from chosen policies to realized outcomes (Harrington (1993), Majumdar and Mukand (2004), Fu and Li (2014), or Willems (2013)).⁴ As in our setting, politicians are trading off the potential welfare and electoral gains when undertaking (and continuing) a policy reform against the reputation loss that a failure would deliver. The focus in our paper is on bringing some predictions of these models to the data. Our analysis is the first to test the implication of reputation concerns for incentives to innovate and learn in the political arena.

Finally, our study is also related to the literature on policy persistence which tackles the question why certain policies persist even if they are known to be a failure. Potential mechanisms stressed in the previous literature are individual-specific uncertainty about the winners and losers of a reform (Fernandez and Rodrik (1991), Jain and Mukand (2003), Ciccone (2004)), a war of attrition between politicians (Alesina and Drazen (1991)), or vested interests benefiting from the status quo (Coate and Morris (1999)). Our focus on the reputation costs of policy reversals as one explanation why policies might persist differs from potential mechanisms highlighted in previous papers. Most importantly, we provide an empirical test whether reputation matters for policy persistence in the context of the U.S. welfare reform.

3 The 1996 Welfare Reform in the United States

3.1 AFDC Program and Welfare Waivers

Since the New Deal, the program ‘Aid to Families with Dependent Children’ (AFDC) provided financial assistance to needy children lacking parental care or support (see

³One exception is Lieberman and Shaw (2000) who analyze the choice of welfare rules as a function of local conditions. Also, Soss et al. (2001) study how an array of factors such as ideology, racial composition or electoral competition, affect the policy choices of states in the TANF program.

⁴Recent papers by Callander (2011a, 2011b), and Callander and Hummel (2014) model policy experimentation with a richer policy space: Their “world” is not restricted to a two-state world and policies in these papers can take more values than the binary policy choice (reform or not) adopted in most of the earlier literature. Callander and Harstad (2015) present a model where heterogeneous districts choose both whether to experiment and the policies to experiment with. Reelection concerns and reputation of the policy makers play no role in their analysis. Instead, we keep a simple setup with two states of the world and a binary decision in order to study the consequences of reputation concerns for policy experimentation and reversals.

Grogger and Karoly (2005), Moffitt (2008), or Bitler and Hoynes (2010)). The program was jointly administered by the federal and state governments though most of the eligibility criteria and provisions were determined at the federal level.⁵ States could chose the level of monthly benefits but had otherwise little room to shape welfare policy.⁶

Starting in the early 1960s, state governments obtained more autonomy to experiment with their own welfare rules. Since 1962, Section 1115 of the Social Security Act allowed the Secretary of Health and Human Services to waive federal AFDC rules and regulations. States could petition the U.S. Department of Health and Human Services (DHHS) to implement other welfare rules in temporary pilot projects. To apply for a waiver, state governments submitted detailed requests which rules and program elements they planned to modify, and what rules and regulations they wanted to implement instead.⁷ The proposed waiver provisions were then reviewed, and sometimes altered, by the federal offices with jurisdiction over the various aspects of each state’s proposal. The Secretary of Health and Human Services made the final decision whether to approve a waiver, request changes or deny it. In most cases, the waiver was approved as is or after some changes.

Welfare waivers were rare until the late 1980s when AFDC caseloads approached almost 4 million families or around 11 million recipients nationwide (see Figure A1 for the evolution of per capita caseloads). Waivers became much more popular under the Reagan years when federal aid for AFDC declined and government funds tightened during the 1989-1992 recession. Between 1987 and 1992, 15 waiver applications in 14 states were approved under the Reagan administration and another 15 applications from 12 states were approved under the Bush administration (see Harvey et al. (2000)). As reform efforts accelerated during the first term of the Clinton administration, the federal government approved 83 waivers (see Figure 1 for the evolution of waiver activity during the 1978-1996 period). In total, all but five states received approval for one or more waivers. Figure 2 shows the distribution of waiver applications across U.S. states between 1978 and 1996 ranging from 0 in Alaska to 10 in Hawaii.

State governors in many states played a key role in initiating welfare waivers. Wisconsin’s governor Tommy Thompson, a leading figure in the welfare reform process, provides a good example. He made welfare reform a top priority in his campaign for governor as early as 1986 (Mead (2004), Kaplan (2000)). When elected in 1987,

⁵A separate program for jobless two-parent families (AFDC-UP) was established in 1961, but single-parent families remained the primary beneficiaries of the AFDC program prior to 1996.

⁶For a family of three with no other income, for instance, monthly benefits in 1995 varied from \$119 in Mississippi to \$720 in Alaska. States could also implement a few eligibility rules like “fit parent” or “suitable home” provisions in order to limit payments to families with unsatisfactory behavior.

⁷Most waivers implemented multiple changes to welfare rules simultaneously (see Crouse (1999) and Koerper (1996) for detailed descriptions of the state waivers). Most popular were experiments to impose stricter work requirements (to increase work participation among recipients), time limits on benefit receipt, sanctions (for benefit units failing to meet work requirements), and family caps (eliminating or reducing the benefit increase if a child is conceived during welfare receipt).

Thompson immediately created a task committee to reform the existing AFDC system. The Thompson administration applied for three waivers in 1988, 1992 and 1993.⁸ In 1993, Governor Thompson used his line item veto to implement the new Wisconsin Works, better known as W-2. With its emphasis on labor market participation, stricter work requirements and harsh sanctions for noncompliance, the W-2 program became a blueprint for the federal welfare reform in 1996. Governors in Delaware, Michigan, Minnesota, Nebraska and Ohio played similar defining roles for the speed and direction of welfare reform (see Weissert (2000) or Winston (2002) for informative case studies).

3.2 The Introduction of TANF

The passage of the Personal Responsibility and Work Opportunity Reconciliation Act (PROWRA) in August 1996 abolished AFDC and replaced it with ‘Temporary Assistance to Needy Families’ (TANF).⁹ The 1996 reform reshaped the landscape of U.S. welfare policy along at least three dimensions: First, the reform decentralized the authority to design welfare programs to the state level. State governments were now free to choose their own welfare rules, however generous or restrictive, bound only by a few federal guidelines. In practice, waiver provisions remained in place until the waiver expired or were replaced by new statewide TANF programs.¹⁰ Because states had to decide on many new policy dimensions – like eligibility rules, work requirements, sanctions or time limits – policy rules have changed even after the adoption of TANF at the state level and do so until today.

The second major element of the reform was the explicit goal to reduce welfare dependency and boost work participation (i.e. workfare instead of welfare). By setting federal time limits to benefit receipt, for example, TANF introduced the idea that government support was a temporary measure rather than a permanent source of income.¹¹ Work requirements became much stricter in the TANF legislation compared to AFDC, and focused on active participation in the labor market rather than training and education. Recipients who failed to comply with the new work requirements faced more rigorous sanctions than under AFDC.

Finally, federal funding after the TANF reform changed from a matching grant into

⁸In 1988, for example, Wisconsin was the first state to make transfer receipt of a household conditional on the school attendance of its teen children.

⁹State governors played a prominent role in pushing and keeping federal welfare reform on the political agenda as well. The National Governors’ Association (NGA) was influential in lobbying for welfare reform at the federal level (see Weaver (2000)). Fourteen governors testified in welfare hearings of the 104th Congress compared to only three state representatives (Winston (2002)).

¹⁰State TANF programs replacing existing AFDC programs were implemented between September 1996 (Massachusetts, Michigan, and Vermont) and as late as January 1998 (California).

¹¹The federal government set a lifetime time limit for benefit receipt of 60 months. States that decided to offer longer time limits (or no time limit at all) have to finance recipients beyond the federal time limit out of state funds. States could also decide to implement shorter lifetime time limits on benefit receipts. We discuss the specific welfare rules that are part of the empirical analysis in more detail in Section 5.

a single block grant. As a consequence the states had to finance rising caseloads or of less restrictive eligibility criteria with their own funds.¹² The federal government also introduced financial penalties, in the form of percentage reductions in the block grant, on states that did not comply with the federally mandated work participation rates of recipients or had excessive caseloads. As a consequence, states had strong incentives to keep their spending levels and welfare caseloads in check in order to avoid any additional burden on the state budget. These incentives are also reflected in the actual numbers: Caseloads, for example, fell nationwide by more than 56% between 1994 and 2000 (see Bitler and Hoynes (2010)).¹³ Despite a common focus on reducing welfare dependency, states differed a lot in the extent and speed of policy changes during both the waiver and the TANF period (see Figures 1-4). Yet, how can we explain that some states experimented more intensively with welfare policies than others? We next provide a theoretical framework to explore the political forces that shaped policy experimentation during the U.S. welfare reform.

4 Theoretical Framework

To study the decision to experiment with (and possibly reverse) welfare policies in the U.S. states, we outline a simple model inspired by Majumdar and Mukand (2004). Our main goal is to derive predictions which we can then take to the data. Given the dominant role played by state governors both during the waiver and the TANF period, we focus on a single political decision-maker who has to decide whether to experiment with an unknown policy or not. To simplify the exposition, all proofs are relegated to the Appendix.

Policies. Consider an elected politician who has an opportunity to enact a reform. The incumbent needs to decide whether to continue with the safe status quo policy a_S or experiment with the new policy a_N . Both policies affect overall welfare which is publicly observable. Different policies are appropriate for different environments; hence, the success of the new policy a_N is contingent on the underlying state of the world. The state of the world may be one of two types denoted by S and N , with a_S and a_N being the appropriate policies for the two environments respectively. If the underlying state of the world is S , then enacting the new policy a_N causes a net loss in welfare. The state N (resp. S) occurs with probability p (resp. $1 - p$).

¹²The size of each state's block grant was fixed at the federal spending level in the 1992-1995 fiscal years. States were further required to contribute substantially to other program components like subsidized child care. Under the old AFDC program, in contrast, a matching grant by the federal government co-financed state and local contributions such that the costs of rising caseloads were shared between the state and the federal governments.

¹³The literature suggests that the booming U.S. economy of the 1990s is responsible for about a third of the caseload reduction; the remainder is likely due to states' incentives to tighten access to welfare benefits (see, for example, Council of Economic Advisors (1999) and Fang and Keane (2004)).

The status quo policy a_S is assumed to deliver a fixed welfare, normalized to 0, independently of the state of the world. No benefits from the new policy occur in state S . In state N , the new policy initiative delivers a benefit Δ . However, this benefit occurs either in the short-term or in the long-term. The benefit realizes in the short term (resp. in the long run) with probability q (resp. $1 - q$).

The total cost of enacting and continuing the new policy initiative is c which is incurred independently of the success of the reform. This cost may represent the additional training of the bureaucracy to effectively administer the new policy or other costs of implementation. The total cost c consists of short- and long-run costs so that $\frac{c}{2}$ is incurred when the reform is implemented and again if the reform is maintained in the long term. If the reform is reverted back to a_S instead, then the long-term cost is not incurred. However, the short-term benefit cannot be kept in the long term if the reform is reversed.

We make the following assumptions:

$$p\Delta - c > 0, \quad (A1)$$

$$pq\Delta - \frac{c}{2}(1 + pq) > 0, \quad (A2)$$

and

$$\beta\Delta - \frac{c}{2} < 0, \quad (A3)$$

where $\beta = \frac{p(1-q)}{1-pq} < p$ is the probability that the state of the world is N conditional on a short-term failure of the reform.

Assumption (A1) implies that a reform that is continued has positive expected welfare gains ex-ante. Assumption (A2) implies that undertaking the reform and reverting back to the status quo in the face of short-term failure has positive expected benefits ex-ante. Finally, assumption (A3) implies that, if the benefit does not realize in the short term, the expected future benefit from continuing with the reform is negative.¹⁴

Timing. The game has three periods, $T = \{0, 1, 2\}$. At the beginning of the game ($T = 0$), the incumbent politician has the opportunity to change the existing policy a_S . She receives a private signal about the appropriateness of the policies, i.e. the state of the world, and then faces the choice of either maintaining the status quo policy a_S or enacting the new policy a_N . If she decides to maintain the status quo, then the ex-post welfare is 0. On the other hand, if the new policy is enacted, a cost $\frac{c}{2}$ is initially incurred. Then, the politician learns about the realization of short-term benefit from the reform at the beginning of $T = 1$. If the new policy turns out to be successful, the politician knows that the state is N , and therefore a_N is the appropriate policy. This implies that the benefit Δ is secured by continuing the reform and incurring the long-term cost $\frac{c}{2}$ at the

¹⁴We implicitly assume that the parameters of the model take values such that assumptions (A1), (A2), and (A3) hold simultaneously. This is the case for a large range of parameter values.

beginning of period $T = 2$. However, in the case where the benefit does not realize in the short term, it is unclear whether the failure is because of bad luck or because the policy is not appropriate. The politician now has to decide whether to continue with the policy initiative a_N or to revert back to the initial status quo a_S . If she reverts back to a_S , ex-post welfare is $-\frac{c}{2}$. If she continues with a_N , the society again incurs a cost $\frac{c}{2}$ and gets the long-run benefit only if the state of the world is N . The ex-post welfare is observed before the next election that takes place at the end of period $T = 2$, corresponding to the end of the incumbent's term in office.

Politicians. The economy has many politicians that differ in their capability to assess the appropriateness of policies for the economy, and can be either of high quality or of low quality. The incumbent politician knows her type. For simplicity, we assume that a high-quality incumbent knows the state of the world. On the other hand, a low-quality politician does not know anything ex ante, i.e. she only knows that the state of the world N is realized with probability p , and that in this state the benefit realizes in period 1 (resp. period 2) with probability q (resp. $1 - q$).

The incumbent politician cares about both welfare of her constituency as well as her own future electoral prospects. The incumbent's objective function is

$$\gamma(\text{Welfare}) + (1 - \gamma)(\text{Probability of Reelection}), \quad (1)$$

where $\gamma \in [0, 1]$ is the relative weight that she puts on welfare.

Finally, at the beginning of the game ($T = 0$) the incumbent has earned a reputation among the electorate represented by $\lambda \in [0, 1]$. We assume that this reputation does not depend on future decisions taken by the incumbent and remains constant until the next election.¹⁵

Voters and Reelection Rule. The electorate consists of a representative voter that chooses either to reelect or throw the incumbent out of office at the end of the term $T = 2$. The voter cares only about ex-post welfare. We assume that the voter does not know that there are high- or low-quality politicians so that he does not think that the incumbent and the (large number of) potential challengers differ in competence. However, the voter may take into account the reputation of the politician when deciding to reelect the incumbent. Assuming the representative voter does not know that politicians differ in abilities simplifies the analysis considerably. Under this assumption, there is no signaling game in the policy choice of the incumbent: Both the challenger and the incumbent are perceived to be identical by the voter except in the case where the incumbent's reputation

¹⁵See Majumdar and Mukand (2004) for an analysis where the reputation of the incumbent changes with the decision to undertake reforms.

is positive.¹⁶

We assume that the representative voter follows a simple reelection rule: Reelect the incumbent with probability 1 if ex-post welfare is positive; never reelect the incumbent if ex-post welfare is negative; and reelect the incumbent with probability λ if ex-post welfare is 0. This reelection rule is in the spirit of agency models where voters punish the incumbent for bad performance (see Persson and Tabellini (2000)).

4.1 Policy Choices of a High-quality Politician

We first analyze the decision of a high-quality incumbent.

Proposition 1. *For any relative welfare weight γ and any reputation λ , a high-quality incumbent always undertakes a reform if and only if the state is N . In addition, a reform undertaken is never reverted back to the status quo.*

A high-quality politician knows the state of the world. If the state of the world is S , since there are no welfare gains and therefore no electoral gains from a reform, she will never experiment. Proposition 1 shows that a high-quality incumbent has her objective fully aligned with the interest of the representative voter: There is no conflict of interest between maximizing voter welfare and her reelection chances.

4.2 Policy Choices of a Low-quality Politician

We now show that the decision-making of a low-quality incumbent may exhibit a conflict between maximizing welfare and reelection chances. In order to disentangle the different forces, we start by developing two polar cases: The one of a purely welfare-maximizing politician ($\gamma = 1$) and the one of a purely office-motivated politician ($\gamma = 0$).

Welfare-maximizing Politician. The following proposition presents the case where the policy choices of the incumbent are driven only by welfare considerations.

Proposition 2. *A low-quality incumbent who only cares about welfare ($\gamma = 1$) always implements the reform, and always reverts back to the status quo if the reform does not deliver the short-term benefit.*

The decision-making policy described in Proposition 2 is the optimal one from a welfare perspective under the veil of ignorance. Indeed, for a low-quality politician who does not know the state of the world, the reform is welfare maximizing from an ex-ante perspective. However, it is welfare-maximizing to revert back to a_S if the reform fails in the short run since the expected future benefit from the reform is then negative.

¹⁶The signaling game where the representative voter makes inferences on quality of the incumbent based on the sequence of policy choices and the realized outcomes is analyzed in Majumdar and Mukand (2004). The predictions they derive are similar to the ones we obtain in the simplified model presented here.

Office-motivated Politician. The second benchmark case presents the policy choice of a purely office-motivated politician.

Proposition 3. *A low-quality incumbent who is purely office-motivated ($\gamma = 0$) implements the reform if and only if*

$$p \geq \lambda, \tag{2}$$

and never reverts back to the status quo when the reform is undertaken.

Proposition 3 shows that it is the reelection motive of the low-quality politician that is responsible for policy persistence in the face of short-term failure. A low-quality politician might therefore not maximize voter welfare for two reasons. First, equation (2) implies that the politician’s initial reputation can conflict with the decision to experiment with profitable reforms. Second, in the face of short-term failure a purely office-motivated politician does not take into account the expected welfare loss of continuing with the reform and “gambles” for reelection with the welfare of the electorate.

Corollary 1. *Purely office-motivated politicians experiment less when they have higher initial reputation among the electorate. In particular, there exists a cut-off $\lambda^* := p$ such that low-quality politicians try the reform if and only if their reputation is below λ^* .*

Corollary 1 characterizes the cut-off λ^* which determines the reputation level that leads to inefficient policy experimentation. Indeed, following Proposition 2, an uninformed politician should always try the reform to maximize ex-ante welfare. However, a low-quality politician with a high enough initial reputation, i.e. above λ^* , will never experiment with the reform. As a consequence, reputation concerns result in too little policy experimentation from a welfare perspective.

Intermediate Politician. We now turn to the general case where the low-quality politician is concerned with both welfare and her reelection chances ($\gamma \in [0, 1]$). We focus on the decision to revert an experiment.¹⁷

Proposition 4. *An incumbent politician with low-quality always reverts back to the status quo in face of a short-term failure of the reform if and only if $\gamma \geq \gamma^*$, where*

$$\gamma^* := \frac{1}{1 + \frac{c}{2\beta} - \Delta}. \tag{3}$$

A low-quality politician who observes the short-term failure of her reform does not know whether the reform will be successful in the long term. This effect pushes her

¹⁷For completeness we also treat the decision to experiment in Appendix A.2. We relegate this case to the Appendix because it does not yield any important insights for our empirical analysis beyond those discussed in Propositions 2 and 3.

to revert back to the status quo in order to increase expected welfare as described in Proposition 2. However, reelection concerns pull her towards sticking with the reform as in Proposition 3. Proposition 4 disentangles these two conflicting forces which affect the decision to revert back to the status quo or not. Equation (3) defines a threshold level γ^* so that for values of relative welfare weights above the threshold, the low-quality incumbent reverts the reform as welfare considerations outweigh reelection considerations.

Corollary 2. *A low-quality incumbent is more likely to revert in the face of short-term failure of the reform when γ increases and Δ decreases.*

Corollary 2 shows how the likelihood of a policy reversal is related to the model parameters which leads to a better alignment of the objectives of the low-quality incumbent and the representative voter, as described in Proposition 2. High-quality politicians never revert an experiment (Proposition 1). In contrast, low-quality politicians make their reversal decision dependent on the welfare gains and costs of the experiment (Corollary 2).

4.3 Empirical Predictions

The propositions above characterize how reelection concerns affect policy experimentation and reversals, and identify four testable implications. The most important empirical prediction is that politicians with higher initial reputation are less likely to experiment. Politicians, especially if they are purely office-motivated, are less likely to reform because they weigh the risky reputation gains from an experiment against their safe initial reputation (Proposition 3 and Corollary 1). The reputation effect should in turn be weaker for politicians with a higher weight on voter welfare. As the experiment increases welfare ex-ante, politicians who are more concerned about welfare are more likely to experiment than politicians who care mostly about their own reelection (Proposition 2 as well as Proposition 5 and Corollary 3 in the Appendix for intermediate cases).

Our second prediction is that experimentation increases with the expected gain from experimentation. Among intermediate politicians, higher potential gains (either because the probability that the state of the world calls for a reform; or because the gains of a successful experiment are high) increase the threshold level for which welfare concerns outweigh reelection concerns (Corollary 3 in the Appendix).

A third prediction is that a low-quality politician may be more or less likely to experiment than a high-quality politician. The high-quality politician experiments when the state of the world is N which occurs with probability p (Proposition 1). The low-quality politician is more likely to experiment if $\gamma = 1$ (Proposition 2) but less likely to experiment if $\gamma = 0$ and $\lambda \geq p$ (Proposition 3). Yet, a low-quality politician with a high weight on voter welfare is more likely to experiment than a low-quality politician who puts a low weight on voter welfare (Corollary 3 in the Appendix).

Our final prediction relates to policy reversals: Low-quality politicians are more likely to revert an experiment than high-quality ones. Low-quality governors do not know whether the experiment is appropriate for the current state of the world; hence, she is more likely to have made a mistake. Also, a low-quality politician is less likely to revert if the potential welfare gains from an experiment are high (Proposition 4 and Corollary 2). In contrast, the initial reputation of the politician has no effect on the likelihood of a reversal. The reason is that a low-quality politician who reverts back to the status quo is voted out of office for sure. A low-quality politician who does not revert is either successful in the long run (in which case she is reelected for sure) or is not successful (in which case she is voted out of office).

We next introduce our data and discuss our empirical strategy to test these predictions on policy experimentation and reversals in the context of the U.S. welfare reform.

5 Data Sources and Empirical Strategy

5.1 Data Sources

To test our predictions, we collect a comprehensive dataset of welfare policy rules, political conditions and socio-economic characteristics for all U.S. states between 1978 and 2010. Here, we introduce our key variables. The appendix provides more details about the data sources and the construction of all variables used in the empirical analysis.

Policy experimentation and reversals. To measure policy experimentation during the AFDC period, we use the number of waivers in each state (from Koerper (1996), Crouse (1999), and Lieberman and Shaw (2000)). Our measure counts waiver applications irrespective of whether the state planned to implement it only in specific counties or the state as a whole. To reduce any bias by the federal review process, we include waiver applications that have been approved and implemented, but also those that were denied by the Secretary of Health and Human Services or withdrawn by the state. Figure 2 shows the distribution of waivers across states. Most states only apply for one waiver within a year while twenty states applied for two waivers within the same year. Our measure of policy experimentation prior to 1996 is then a simple count variable of the number of waivers a state applied for in a year. If a state did not apply for a waiver, the count variable is set to zero. On average, states applied for 0.2 waivers per year during the 1978-1996 period.

Identifying a policy experiment after the federal welfare reform in 1996 is more challenging. We obtain information on welfare policy rules in each state and year from the Welfare Rules Database of the Urban Institute (Urban Institute (2015)) which provides the most comprehensive and up-to-date information on state welfare rules for the post-1996 period. The database contains literally hundreds of rules about eligibility, benefit

calculations, sanctions and time limits and many other requirements. For our analysis, we focus on a subset of policy rules that were at the center of the public and academic debate during the reform period (see the discussion in Bitler and Hoynes (2010); Fang and Keane (2004); or Grogger and Karoly (2005) for examples): (1) Whether the state adopted a family cap, which limits the additional benefits if a child was conceived and born while the parent is on welfare. (2) The rules imposed for time limits on benefit receipt. (3) The type of work requirements adults had to fulfill in order to remain eligible for benefit receipt. (4) The primary sanctions imposed if a household did not fulfill the work requirements. The data appendix and Table A1 provide more details about each individual policy rule.

To illustrate how we measure an experiment or reversal, take the example of family caps which did not exist under the old AFDC program. A family cap limits the amount of additional benefit a recipient household may receive if an additional child is born in the household while on welfare. Between 1996 and 2010, twenty-four states (among them, California, Florida and Illinois) adopted a family cap, eight of these twenty-four states (among them, Illinois and Maryland) later abolished the family cap previously adopted. We then code a policy experiment if a state first adopts a family cap in any year after 1996 and zero if no family cap was adopted; we code a policy reversal if a state (that had adopted a family cap until 1996 or later) abolishes it in any year between 1997 and 2010. We code policy reversals of other welfare rules in a similar way: A tightening of a policy rule in a specific year is coded as an experiment, abolishing that restriction in a later year is coded as a policy reversal.¹⁸ Figure 3 shows that many policy experiments occur shortly after the PRWORA reform, while policy reversals are rare throughout the whole period. Figure 4 demonstrates that experimentation again varied a lot across states; it was much more common in some states (Pennsylvania, for instance) than others (like New York).

Electoral incentives and potential gains. Our main independent variables are the governor’s initial reputation (an empirical counterpart to λ), the importance of reelection concerns (a proxy for $1 - \gamma$); some measure indicating the potential gains

¹⁸Note that we can measure the adoption and abolishing of specific policy rules unambiguously only when they are covered by the Welfare Rules Database of the Urban Institute. Changes in policy rules during the waiver period are not captured systematically in the database because the information contained in the waiver application cannot be merged to the Welfare Rules Database in a consistent manner. The coverage of rules is most complete and reliable after 1996. Hence, if a state both adopted and abolished a family cap prior to 1996, we will not capture this policy change in our measure of policy reversal. Anecdotal evidence however suggests that policy reversals were rare before 1996 as many waivers only expired after the federal TANF reform. Similarly, we might have a bias in our post-1996 policy experimentation measure as well. Suppose a state experiments with a policy rule, say a family cap, using a waiver, but did not have a family cap in 1996 (as coded in the Welfare Rules Database). If the waiver provisions are still in place in 1996, any changes will be captured by our experimentation measure; if in contrast, the waiver expired before 1996 and the state again adopted a family cap after 1996, we would code it as an experiment (even though the state has tried this experiment before in the waiver period).

from experimentation (a proxy for Δ); and finally, a proxy for the governor's knowledge whether the state of the world calls for an experiment. All of these concepts are difficult to measure as they are ultimately unobservable to the researcher.

We use the vote margin in the past gubernatorial race to proxy the governor's initial reputation. Governors who were elected with a larger majority are also more likely to be reelected conditional on the state of the world and the type of challenger. To rule out that the past vote margin just captures electoral competition in a state, we further control whether the party of governor switched during the past electoral races.

To measure reelection concerns among governors, we use information on whether a state governor faces a binding term limit in her current term in office, i.e. whether she is a lame duck. Lame ducks might still have some electoral incentives, for example, because they plan to run for another office at a higher level. All we need for our analysis is that lame ducks care somewhat less about their electoral prospects than governors who can and want to remain in office in their current state.

The most difficult task is to find a proxy for the potential gains of a policy experiment. An important motivation for state governors to apply for a waiver was to reduce caseloads and hence, welfare spending. The incentive to reduce state spending became even stronger during the TANF period because states have to cover all welfare expenditures exceeding the federal block grant.¹⁹ We therefore use a state's AFDC/TANF spending (lagged by two years) to proxy potential gains from policy experimentation. The idea here is that a state who spends a lot on welfare is more likely to search for policy rules to effectively reduce its welfare burden because of financial constraints; or, because voters may benefit in some other way from an effective reduction in welfare spending.

Finally, we require a measure of governors' competence to identify when an experiment is called for. We use the years a governor has spent in all political offices before being elected as governor (see also Smart and Sturm (2013) for a similar approach). The accumulated number of years in prior political offices is a good indicator of the governor's ability to recognize the state of the world (and when an experiment is called for) either if elections select the most able candidate; or if candidates acquire competence in political decision-making under uncertainty over time (see, for example, Alt et al. (2011) for recent evidence). While voters may in principle observe her tenure as a governor, they are unlikely to know exactly how much political experience a governor has accumulated prior to being elected.²⁰

¹⁹The block grant was also cut if states did not reach the federally mandated employment rates among benefit recipients, for example.

²⁰A few recent studies use formal educational qualifications as a proxy for quality instead (see Galasso and Nannicini (2011) and Gagliarducci and Nannicini (2013) for Italian politicians or Kotakorpi and Poutvaara (2011) for Finnish members of parliament). In our analysis, educational qualifications were never found to be statistically or economically relevant for policy experimentation. One possible explanation is that most governors in our sample are highly skilled: Only 6% do not have a college degree or more, while almost 50% have a master or other postgraduate degrees. In contrast, the average local politician in Italy only has a high-school degree; in the Finnish parliament, about 20-30% of its members

Control variables. To control for other potential influences on the decisions to experiment and reverse a policy, we collect information on the political and socio-economic characteristics in each state. Our first set of control variables consists of several measures of state ideology. We obtain the governor’s party and a measure of governor ideology (based on Berry et al. (1998)) which takes into account that governors belonging to the same party might have different ideologies. The ideology measure varies between 0 and 100 where larger values represent a more liberal attitude. We also use Berry’s measure of citizen ideology as well as the voter support for the Democratic candidate in the last Presidential election as additional measures of state ideology.

To control for legislative influences, we use measures of party composition and political polarization. Polarization is measured as the Democratic seat share in the lower (upper) chamber calculated as deviation from 50%. We further code whether the government in the state is politically divided between Democrats and Republicans; and a measure of competitiveness in state legislative elections using the Holbrook-Van Dunk index (see Holbrook and Van Dunk (1993)).

States might also learn from or imitate other states. To analyze these cross-state spillover effects, we define geographic neighbors (states sharing a boundary), states similar in population size (based on population in 1978) and ideological neighbors in the same Census region (with a similar voting record in the 1978 presidential election).

Finally, policy experiments might also be influenced by the local demand for welfare. To control for the influence of the local population, we collect information on personal income per capita, the size and composition of the population, the share of immigrants, the share of unmarried births and income inequality (measured as household income at the 90th over the 10th percentile). Table A2 contains summary statistics of all variables used in our analysis separately for the waiver (1978-1996) and TANF period (1997-2010).

5.2 Empirical Strategy

For the empirical analysis, we use a simple difference-in-difference approach. To test our first prediction, we estimate variants of the following model:

$$Experiment_{st} = \alpha_1 Reput_{st} + \pi_1 LD_{st} * Reput_{st} + \phi_1 LD_{st} + \mu_1 Gains_{st} + t_t + \theta_s + \varepsilon_{1st}, \quad (4)$$

where $Reput_{st}$ measures the governor’s initial reputation, LD_{st} is an indicator whether the governor faces a binding term limit and $Gains_{st}$ measures the potential gains from experimentation. θ_s denotes state fixed effects and t_t year fixed effects. We expect policy experimentation to decrease with the incumbent’s reputation, i.e. $\alpha_1 < 0$. A politician with high initial reputation will be reluctant to engage in risky policy experiments, while a

have at least a masters degree. Hence, formal educational credentials might be more a prerequisite for becoming governor rather than a signal of competence. Furthermore, voters are probably more aware of the educational credentials than of the governor’s prior political experience.

politician with low reputation is more likely to experiment hoping to boost her reelection chances. In addition, we expect the reputation effect to be weaker for governors who cannot be reelected and hence worry less about their reputation, so $\pi_1 > 0$. Our second prediction is that the likelihood of an experiment increases with the potential gains from experimentation, i.e. $\mu_1 > 0$.

One might worry that reverse causality could bias our estimates in equation (4). We address these concerns in two ways. First, we lag the explanatory variables (initial reputation and potential gains) to ensure that contemporaneous shocks do not influence both experimentation and the independent variables. Obtaining the status of a lame duck in turn is largely determined by the term limit provisions in each state. Second, we think that a potential endogeneity bias should work in the opposite direction than the effects we predict from our model in Section 4. We expect that higher AFDC or TANF spending in the past increases the likelihood of experimentation; in contrast, experimentation today should reduce future spending as a tightening of rules and eligibility criteria reduces future caseloads or the number of beneficiaries. Similarly, if a governor campaigns on implementing welfare experiments, we would expect that the decision to experiment with welfare rules will boost her future vote margin. Our model in contrast predicts that the past vote margin reduces experimentation because of the potentially reputational costs of failure. In both scenarios, our estimates for the potential gains from experimentation and the reputation effect should be a lower bound (in absolute terms).

Whether governors with less experience are more or less likely to experiment than a governor with a lot of experience is a priori unclear. However, we should observe that governors with little experience are more likely to experiment if they are less concerned with reelection. To test this prediction, we estimate variants of the following model:

$$Experiment_{st} = \alpha_2 LExp_{st} + \pi_2 LExp_{st} * LD_{st} + \phi_2 LD_{st} + \mu_2 Gains_{st} + t_t + \theta_s + \varepsilon_{2st}, \quad (5)$$

where $LExp_{st}$ is equal to one if the governor has less political experience (below the median) and zero otherwise. All other variables are defined as before. We expect politicians who have little experience but cannot be reelected to experiment more, $\pi_2 > 0$, while the sign of α_2 is theoretically ambiguous. Again, we should observe that higher gains make experimentation more likely, i.e. $\mu_2 > 0$.

Could the coefficient on political experience in equation (5) be affected by endogeneity? Recall that we calculate a governor's experience as the number of years in political office prior to being elected as governor. Hence, political experience can be considered predetermined (conditional on being elected as governor in the first place).²¹

Our final prediction covers policy reversals which we implement using variants of the

²¹Reverse causality would be an issue if welfare reform and the option to experiment change the selection of candidates into gubernatorial races with respect to their prior political experience. We think this is a rather unlikely scenario.

following model:

$$Reversal_{st} = \alpha_3 LExp_{st} + \pi_3 Gains_{st} * LExp_{st} + \mu_3 Gains_{st} + t_t + \theta_s + \varepsilon_{3st}, \quad (6)$$

where all variables are defined as before. Governors with less experience are more likely to revert a policy that failed; hence, we expect that $\alpha_3 > 0$. Yet, the higher the gains from a successful experiment, the less likely a policy reversal becomes; so, $\pi_3 > 0$. We now turn to our empirical results.

6 Empirical Results

6.1 Experimentation during the Waiver Period

Table 1 shows the results for policy experimentation during the 1978-1996 period. The dependent variable here is the number of waiver applications under AFDC. All specifications control for state and year fixed effects as well as governor age to adjust for general experience. In line with our model, we find that governors with high initial reputation among voters experiment less in the waiver period (see equation (4)). The relationship becomes slightly stronger when we control for swing states by including an indicator whether the governor has changed over the past two electoral cycles. Hence, the reputation effect cannot be explained by the fact that swing states with tougher electoral competition experiment less. Instead, policy experiments in a state are less likely because the current governor has more to lose (in terms of reelection chances) if her reputation is high and the experiment turns out to be a failure.

The second specification (reported in columns (3) and (4)) shows that the reputation effect is indeed much stronger for governors who are not lame ducks. The effect is economically sizable. An increase in the past vote margin by an interquartile range (75p-25p) reduces policy experiments among reelectable governors between 40% and 42%.²² For lame ducks in contrast, the sum of the coefficient on the vote margin and the interaction effect of vote margin and lame duck is very small for all specifications. The last two columns show that the reputation effect is unchanged if we also control for the potential gain from experimentation. As predicted by the model, higher potential gains (as measured by past AFDC spending) are associated with more experimentation. An increase in welfare spending by an interquartile range increases experimentation by 50%.²³

Table 2 tests our third prediction (see equation (5)) which links policy experimentation to the governor's experience. In line with the model's predictions, the first two columns

²²The interquartile range of the past vote margin in our time period is 9.1 percentage points, while the mean number of waivers per year is 0.2. Hence, $(-0.0093 * 9.1) / 0.2 = -0.4232$.

²³The interquartile range of welfare spending during the 1978-1996 period is 250 (US\$ million). Hence, we get $250 * 0.0004 / 0.2 = 0.50$.

confirm that the prior political experience of the governor alone is not systematically associated with a policy experiment. Governors with little experience who cannot be reelected however, indeed experiment more (see columns (3) and (4) of Table 3). The size of the effect is again economically significant: Governors who have little experience and cannot be reelected are 43% more likely to experiment than other governors.²⁴

The last two columns test all three predictions jointly to check whether each factor has a separate influence on the decision to implement a policy experiment. The results confirm that a high initial reputation (the first prediction), higher potential gains (the second prediction) and lame ducks with little experience (the third prediction) have a statistically and economically significant influence on the decision to apply for a welfare waiver. The size of the effects in the joint model is similar to those in the previous specifications suggesting that multicollinearity between the variables is not an issue. While the dependent variable in Table 1 and 2 is a count variable, we find very similar results if we use a binary indicator for a waiver experiment instead (see Table A3 in the appendix for the results).

6.2 Experimentation and Reversal during the TANF Period

Our results thus far indicate that reputation concerns play an important role for applying for a welfare waiver during the AFDC period. Do we observe a similar pattern after the 1996 reform when states could (and had to) choose their own welfare rules?

Table 3 reruns the same models where the dependent variable is now the number of policy experiments during the TANF period. For reputation, we find very similar results as for the waiver period. A higher initial reputation reduces the probability to engage in risky experiments (see column (1) of Table 4); and the negative effect of initial reputation is especially strong among governors who might and can be reelected (see column (2) of Table 4). We also find some support for the prediction that higher TANF spending encourages experimentation with new policy rules (see column (3) in Table 4).

Economically, both reputation and potential gains have, however, a much weaker effect on policy experimentation during the TANF period. An increase in the vote margin by an interquartile range reduces experimentation during the TANF period by only 18%. An increase in potential gains by an interquartile range in turn increases experimentation by about 22%.²⁵

We find generally weaker support for the prediction that governors who have little experience and cannot be reelected are more likely to experiment. While the coefficient on the interaction term (in column (4) of Table 4) is positive, it does not reach statistical

²⁴Both competence and lame duck are binary indicators. Hence, summing the main effects and the interaction effect and dividing by the mean number of waivers, we get: $(-0.09+0.1968-0.0203)/0.2=0.4325$.

²⁵During the TANF period, the interquartile range of the past vote margin is 11.31 percentage points. The mean number of experiments per year is 0.497. Hence, $(-0.0078*11.3)/0.497=-0.178$. Similarly, the interquartile range of TANF spending (in US\$ million) is 356.76. Hence, $0.003*356.76/0.497=0.215$.

significance at conventional levels. Similarly, once we include all controls simultaneously only the reputation effect remains statistically and economically important (see column (5) of Table 4).²⁶

One possible explanation for the weaker effects on policy experimentation during the TANF period is that changes in policy rules were no longer a salient issue for reelection. In our theoretical model, experimentation is decisive whether a politician is reelected or not. After the federal PRWORA reform in 1996, welfare policy in the U.S. states might have turned from a highly controversial political topic to a more practical issue of managing and administrating welfare cases which created little voter attention during election campaigns.

For the TANF period, we can also test our last prediction about policy reversals (see equation (6)). An important caveat is that in the model, the decision to revert a policy experiment is only taken when the experiment does not deliver short-term benefits. Unfortunately, we do not observe whether the experiment failed in the short term or not. Our reversal measure therefore contains substantial measurement error which reduces the precision of our estimates. To reduce this concern, we focus here on substantial changes in welfare rules to define a policy reversal. The reason is that substantial adjustments in welfare rules are more likely to imply a failure of existing rules than minor adjustments which might simply reflect administrative changes or convergence to a perceived optimum.²⁷ Nevertheless, we view the results for reversals as being more suggestive in nature.

Table 4 reports our estimates where the dependent variable is now the number of policy reversals in a state and year.²⁸ Governors with little prior experience are more likely to revert an experiment. This result is in line with our model as low-quality governors are more likely to have made a mistake (by implementing an experiment in the wrong state of the world) and hence, are more likely to revert a policy. In the second specification, we also include the initial reputation of the governor which, according to our model, should not play a role for the decision to reverse a policy experiment. And indeed, the correlation between initial reputation and a policy reversal is very weak (relative to the effect on policy experimentation in Table 2, for example) and never close

²⁶We also construct two alternative experimentation measures: The first one restricts experiments to the first change in a rule but ignores subsequent rule changes in the same direction. The second alternative uses an extended set of welfare rules which also includes earnings disregards for benefit calculation; whether a state requires households to look for a job prior to applying for benefits (mandatory job search); and whether a state offers temporary payments to families if they do not apply for welfare benefits (so-called diversion payments). We obtain similar results if we focus on the first change in each welfare rule during the TANF period; in contrast, we find few effects for the broader set of welfare rules which were much less subject to public debate than the rules analyzed in the main tables.

²⁷For work requirements, for example, we code a policy reversal as one if the hours requirement changes by at least 30%. By focusing on substantial policy changes, we hope to reduce measurement error from small adjustments of a policy rule.

²⁸Since reversals are rare (see Figure 3), we get the same result if we use a binary indicator for a policy reversal instead.

to statistically significant. The fact that initial reputation has little relationship with the reversal decision suggests that voters do not seem to update their opinion about a politician based on the observed policy choices like an experiment. The absence of updating could be an indication that voters may not be able to fully observe or evaluate the politicians' actions and the policy's consequences within a relatively short period of time.²⁹

A policy reversal among governors with little experience is also more likely the lower the potential gains from experimentation are. Columns (3) and (4) also find support for this prediction. If the potential gains from experimentation are high, governors with little experience prefer to gamble: They stick to a policy experiment and hope that the experiment turns out to be successful in the second period. Together, the two factors can explain a substantial share of the policy reversals observed during the TANF period.³⁰

6.3 Robustness Analysis

Overall, our findings provide strong evidence that reputation concerns influence the decision to experiment with policy rules during the waiver period and after the federal reform in 1996. Yet, our analysis has so far abstracted from other potential influences on policy innovations like ideological differences between governors, the role of the state legislature or the local demand for welfare services. We next discuss and present a number of robustness tests to support our main finding that reputation concerns continue to affect the decision to experiment even conditional on these additional influences.

Ideology. A common perception of the welfare reform, and certainly the federal PRWORA reform, is that it was fueled by a conservative agenda to curb welfare spending. Anecdotal evidence seems to support this notion: Republicans were typically more in favor of the decision to abolish welfare and encourage workfare. How would ideological differences about the best welfare policy affect our analysis? In our model in Section 4, voters and politicians agree on which policy is best given the state of the world which is obviously a strong assumption. We can extend the model to allow for some ideological

²⁹The result that initial reputation does not matter for the decision to revert a policy differs from the prediction in the full signaling game analyzed in Majumdar and Mukand (2004). There, voters update their belief about the politician's quality after observing policy choices and their outcomes. In that case, the initial reputation also has an effect on reversal decisions.

³⁰One would also expect that a policy reversal reduces the reelection chances of the governor. Since welfare is negative if a policy experiment is reversed, the reelection probability of a governor who reverts is always zero in our model. Unfortunately, it is quite difficult to test the link between policy reversal and reelection with our data for two reasons. First, we run into small sample issues as we have to exclude lame ducks and governors who do not run in the next election. Second, a policy reversal may only be observed if an experiment failed in the short term; but failure is not observed by the researcher. In the data, we do find a negative correlation between a policy reversal and reelection for governors who are not lame ducks but the relationship is statistically not very robust due to small sample size and measurement error.

differences between politicians.³¹ One simple way is to assume that Republicans assign a higher probability that welfare reform is the right thing to do, i.e. that the state of the world calls for a reform. Republicans who do not observe the state of the world believe that the state of the world N (where an experiment is welfare-enhancing) occurs with probability $p_R > p$. In contrast, Democrats who do not observe the state of the world continue to have the belief that the state of the world being N is p .³² Since the representative voter cares only about ex-post welfare, the reelection rule remains unaffected by the bias of Republican governors. In such an environment, it is easy to show that Republicans are more likely to experiment with the reform (which increases expected welfare from Proposition 2) and to persist in the face of short-term failure (which decreases expected welfare from Propositions 3 and 4) relative to a Democratic governor.³³

Table 5 reruns the full specification with all controls from Table 2, but adds controls for political ideology as well. Surprisingly, Republican governors are not statistically significantly more likely to experiment than Democratic governors. We find the same result if we use Berry’s measure of party ideology (see Berry et al. (1998)) to account for ideological differences within parties across states (column (3)). The absence of an effect of ideology is consistent with other studies that fail to find any effect of ideology for redistributive politics (see e.g. Aidt and Jensen (2009) for the introduction of the personal income tax; or Bardhan and Mookherjee (2010) for land reform). Republican lame ducks are less likely to experiment (see specification (2)) but the coefficient is again not statistically significant. These results provide little support for the idea that lame ducks, in the absence of reelection motives, follow their party ideology; if anything, they seem to go against their own party (which might be closer to their own bliss point). In columns (4) and (5), we test whether policy experimentation is influenced by voter ideology (measured by Berry’s voter ideology measure and the Democratic vote share in the last presidential election) instead. State-level shifts in voter ideology are, however, not systematically related to the decision to experiment. The final specification shows that including all ideology controls simultaneously does not change our conclusion that reelection concerns are a significant determinant of policy innovations during the U.S. welfare reform.

Legislature. Our main empirical analysis also restricts attention to the governor as the relevant decision-maker. As we argue above, state governors played a prominent role during the waiver period and the passage of PRWORA. Yet, the legislature is certainly

³¹Politicians might also boost their reputation by proposing and implementing policies that are on the other side of their ideological spectrum, see Cukierman and Tommasi (1998) and Moen and Riis (2010).

³²The analysis yields the same prediction if Democrats underestimate the probability of the state being N relative to Republicans instead.

³³To illustrate, take the example where $\gamma = 0$. Then, from Proposition 3, a politician experiments whenever $p \geq \lambda$. Since λ is not correlated with the party of the governor in the model, the inequality is more likely to be satisfied for a Republican. As a consequence, she experiments more.

also an important player in welfare policy-making. Veto players in the legislative process, for example, might delay or reduce policy experimentation while more competition in the legislature or between governor and legislature possibly increases it. Previous evidence suggests that party composition and tightness of legislative elections seem to have an effect on welfare spending under AFDC (e.g. Barrilleaux et al. (2002)).

In Table 6, we add a number of controls for the legislature to our baseline specification in Table 2: The party composition in the lower and upper house; whether the government is divided; the degree of polarization in the parliament (often seen as an indicator for stiff political competition); and finally, the Holbrook and Van Dunk measure of competitiveness. None of the controls for the legislature has a statistically significant influence on policy experimentation. Across different sets of controls for the legislature, most coefficients remain statistically significant though somewhat smaller in magnitude than the baseline in Table 2. The past vote margin loses statistical significance in two specifications: When we control for polarization in the state parliament (column (3)) and once we include all legislative controls simultaneously (column (5)). In both cases however, the coefficient and standard error do not suggest that multicollinearity between the regressors is an issue. The effect is just borderline below the 10% significance level.

Spillover Effects. Our theoretical and empirical analysis also abstracts from spillover effects across states. If potential welfare recipients are geographically mobile, states may engage in welfare competition. In that case, experimentation in one state is influenced by policy choices in neighboring states (see Brueckner (2000), Figlio et al. (1999), Saavedra (2000), or Wheaton (2000)). Spillovers could also arise if policy choices in neighboring states provide information for voters (as in a model of yardstick competition (Besley and Case (1995b)) or politicians. In fact, a central insight of the literature on laboratory federalism is that decentralization promotes learning and spillover of knowledge among decision-makers within a federal system (see Oates (1999) for a survey; Rose-Ackerman (1980); and Strumpf (2002)).

To test for such spillover effects, we create three different measures of learning from neighboring states. Our first measure calculates the share of geographic neighbors who implemented a policy experiment in prior years. Alternatively, we define neighbors as states with similar population size and create their mean propensity to experiment in previous years. Finally, we also consider spillovers from states that are located within the same Census region and ideologically similar (as measured by voter support in Presidential elections). Table 7 shows that there is little systematic statistical relationship between policy experiments in a state and experimentation in neighboring states. The big exception is that more experimentation in states that are ideologically similar and within the same Census region boosts experimentation which points to some policy imitation. These cross-state spillovers between ideologically and geographically similar neighbors also reduce the impact of the past vote margin and the governor’s experience on policy

experimentation. Such a pattern is to be expected, for example, if states learn from similar neighbors whether an experiment is the right thing to do in the current situation (in the model, learning from neighbors would boost p , the probability that the current state of the world calls for a reform). A higher p makes not only experimentation more likely, but also reduces the negative effect of the vote margin on experimentation.

How does the modest evidence for cross-state spillovers square with the rich case study evidence that certain states and governors (like Tommy Thompson in Wisconsin) played a decisive role in the welfare reform process? The key observation here is that cross-state spillovers are important if we do not control for aggregate waiver activity (through year effects). One possible interpretation is that state governors learn from and possibly imitate neighboring states when there is a lot of experimentation overall. Yet, once we control for the aggregate activity pattern of experimentation, spillovers seem to play a less prominent role.

Demand for Welfare Services. We have so far said little about the demand for welfare provisions. One explicit goal of the 1996 welfare reform was to decentralize decision-making to the state level. An important benefit of decentralization is that policies are, under certain conditions, closer to the preferences of the electorate. To test for the influence of the demand side, we add sequentially controls for important socio-economic characteristics like the unemployment rate, income per capita, population size and its composition by age and race, the immigrant share and share of unmarried births as well as income inequality (measured by the 90/10 ratio) to our specification. The last specification (in column (5)) adds all socio-demographic controls simultaneously. The set of demographics controls we include are similar to those commonly used in studies of policy innovations (see, for example, Walker (1969), Berry and Berry (1992), or Boehmke and Skinner (2012)). Table 9 shows that states with a larger immigrant population and more income inequality experiment less.³⁴ We do not find any relationship between the decision to experiment and general economic conditions, population composition or share of unmarried births.

Overall, our robustness analysis suggests that reelection concerns among governors matter for innovations in the public sector even after considering ideological differences, the legislature, spillover effects across states and the local demand for welfare.

7 Concluding Remarks

This article studies which political forces influence policy experimentation and learning from past experiments. Our empirical setting is the U.S. welfare reform in 1996, the most

³⁴PRWORA denied federal welfare benefits to most legal immigrants during their first five years of U.S. residence and placed other restrictions on legal immigrants' eligibility for benefits. these provisions were, however, not yet in place during the waiver period (analyzed in Table 8) and therefore cannot explain the negative correlation observed in the data.

important shift in welfare policy since the New Deal. To motivate our empirical analysis, we outline a political agency model which shows how reelection concerns reduce incentives to engage in policy experimentation and possibly to reverse experiments later on. To test our predictions, we use a new data set of welfare policy experiments and reversals in the U.S. states between 1978 and 2010. We complement our data with detailed information on the governor and legislature as well as additional socio-economic characteristics in each state.

Our predictions confirm that reelection reduces the probability of experimentation both during the waiver period before 1996 and, to a lesser extent, for the period after 1996 when states decided on their own welfare rules. In line with our theoretical considerations, governors who face binding time limits (“lame ducks”) behave systematically different from governors who face reelections and hence worry more about their political reputation. Specifically, lame ducks are more likely to implement an experiment that is welfare-enhancing even when their reputation among the electorate is high. Hence, our findings suggest that mitigating reelection concerns may have benefits when both voters and politicians are uncertain about the best policy. Strong reputation concerns make politicians reluctant to experiment with risky policies and also reduce the likelihood that politicians reverse failed experiments.

References

- Aidt, T. and P. Jensen (2009). The taxman tools up: An event history study of the introduction of the personal income tax. *Journal of Public Economics* 93, 160–175.
- Alesina, A. and A. Drazen (1991). Why are stabilizations delayed? *American Economic Review* 81(5), 1170–1188.
- Allen, R. and J. Clark (1981). State policy adoption and innovation: Lobbying and education. *State and Local Government Review* 13(1), 18–25.
- Alt, J., E. B. de Mesquita, and S. Rose (2011). Disentangling accountability and competence in elections: Evidence from U.S. term limits. *Journal of Politics* 73(1), 171–186.
- Bardhan, P. and D. Mookherjee (2010). Determinants of redistributive politics: An empirical analysis of land reforms in West Bengal, India. *American Economic Review* 100(4), 1572 – 1600.
- Barrilleaux, C., T. Holbrook, and L. Langer (2002). Electoral competition, legislative balance, and American state welfare policy. *American Political Science Review* 46(2), 415–427.
- Berry, F. S. and W. D. Berry (1990). State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review* 84(2), 395–415.
- Berry, F. S. and W. D. Berry (1992). Tax innovation in the states: Capitalizing on political opportunity. *American Journal of Political Science* 36(3), 715–742.
- Berry, F. S. and W. D. Berry (2007). Innovation and diffusion models in policy research.

- In P. A. Sabatier (Ed.), *Theories of the Policy Process*, second edition, pp. 223–260. Boulder, CO: Westview.
- Berry, W. D., E. J. Ringquist, R. C. Fording, and R. L. Hanson (1998). Measuring citizen and government ideology in the American States, 1960-93. *American Journal of Political Science* 42(1), 327–348.
- Besley, T. and A. Case (1995a). Does electoral accountability affect economic policy choices? Evidence from gubernatorial term limits. *Quarterly Journal of Economics* 110(3), 769–798.
- Besley, T. and A. Case (1995b). Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition. *American Economic Review* 85(1), 25–45.
- Besley, T. and A. Case (2003). Political institutions and policy choices: Evidence from the United States. *Journal of Economic Literature* 41(1), 7–73.
- Bitler, M. and H. Hoynes (2010). The state of the safety net in the post-welfare reform era. *Brookings Papers on Economic Activity*, 71–127.
- Blank, R. M. (2002). Evaluating welfare reform in the United States. *Journal of Economic Literature* 40, 1105–66.
- Boehmke, F. J. and P. Skinner (2012). State policy innovativeness revisited. *State Politics and Policy Quarterly* 12(3), 303–329.
- Brueckner, J. K. (2000). Welfare reform and the race to the bottom: Theory and evidence. *Southern Economic Journal* 66(3), 505 – 525.
- Callander, S. (2011a). Searching and learning by trial and error. *American Economic Review* 101(6), 2277–2308.
- Callander, S. (2011b). Searching for good policies. *American Political Science Review* 105(04), 643–662.
- Callander, S. and B. Harstad (2015). Experimentation in federal systems. *Quarterly Journal of Economics*, forthcoming.
- Callander, S. and P. Hummel (2014). Preemptive policy experimentation. *Econometrica* 82(4), 1509–1528.
- Center for Economic and Policy Research (2012). *Current Population Survey ORG Uniform Extracts*, Volume Version 1.7. Washington, D.C.
- Ciccone, A. (2004). Resistance to reform: Status quo bias in the presence of individual-specific uncertainty: Comment. *American Economic Review* 94(3), 785–795.
- Cnudde, C. F. and D. J. McCrone (1969). Party competition and welfare policies in the American States. *American Political Science Review* 63, 858–66.
- Coate, S. and S. Morris (1999). Policy persistence. *American Economic Review* 89(5), 1327–1336.
- Council of Economic Advisors (1999). The effects of welfare policy and economic expansion on welfare caseloads: an update. Technical report, Executive Office of the President of the United States.
- Council of State Governments (2012). *Book of the States*. Washington D.C.: Council of State Governments.
- Crouse, G. (1999). Implementation of major changes to welfare policies, 1992-1998. http://www.aspe.hhs.gov/hsp/Waiver-Policies99/policy_CEA.htm (accessed on February 7, 2014).

- Cukierman, A. and M. Tommasi (1998). When does it take a Nixon to go to China? *American Economic Review* 88(1), 180–197.
- Fang, H. and M. P. Keane (2004). Assessing the impact of Welfare Reform on single mothers. *Brookings Papers on Economic Activity* 1, 1–95.
- Fernandez, R. and D. Rodrik (1991). Resistance to reform: Status quo bias in the presence of individual- specific uncertainty. *American Economic Review* 81(5), 1146–1155.
- Ferraz, C. and F. Finan (2011). Electoral accountability and corruption in local governments: evidence from audit reports. *American Economic Review* 101(4), 1274–1311.
- Figlio, D., V. Kolpin, and W. Reid (1999). Do states play welfare games? *Journal of Urban Economics* 46, 237–254.
- Fu, Q. and M. Li (2014). Reputation-concerned policy makers and institutional status quo bias. *Journal of Public Economics* 110, 15 – 25.
- Gagliarducci, S. and T. Nannicini (2013). Do better paid politicians perform better? Disentangling incentives from selection. *Journal of the European Economic Association* 11(2), 369–398.
- Galasso, V. and T. Nannicini (2011). Competing on good politicians. *American Political Science Review* 105(01), 79–99.
- Grogger, J. and L. A. Karoly (2005). *Welfare Reform: Effects of a Decade of Change*. Harvard University Press.
- Harrington, J. (1993). Economic policy, economic performance, and elections. *American Economic Review* 83(1), 27–42.
- Harvey, C., M. J. Camasso, and R. Jagannathan (2000). Evaluating welfare reform waivers under Section 1115. *Journal of Economic Perspectives* 14, 165–188.
- Haskins, R. (2006). *Work over Welfare: The Inside Story of the 1996 Welfare Reform Law*. Brookings Institution Press.
- Holbrook, T. and E. Van Dunk (1993). Electoral competition in the American States. *American Political Science Review* 87(4), 955–962.
- Holmström, B. (1982). Managerial incentive problems - A dynamic perspective. In *Essays in Economics and Management in Honor of Lars Wahlbeck*. Swedish School of Economics.
- Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies* 66(1), 169–182.
- Jain, S. and S. W. Mukand (2003). Redistributive promises and the adoption of economic reform. *American Economic Review* 93(1), 256–264.
- Kaplan, T. (2000). Wisconsin works. In S. F. Liebschutz (Ed.), *Managing Welfare Reform in Five States: The Challenge of Devolution*. Rockefeller Institute.
- Karch, A. (2007). *Democratic Laboratories: Policy Diffusion among the American States*. University of Michigan Press.
- Klarner, C. (2003). Measurement of partisan balance of state government. *State Politics and Policy Quarterly* 3, 310–319.
- Koerper, K. (1996). Weekly tracking update - welfare reform: Section 1115 Waiver Activity. Office of Planning, Research and Evaluation.
- Kotakorpi, K. and P. Poutvaara (2011). Pay for politicians and candidate selection: An empirical analysis. *Journal of Public Economics* 95(7–8), 877 – 885.

- Leip, D. (2012). Dave Leip's Atlas of U.S. Presidential Elections. <http://www.uselectionatlas.org>. (accessed on February 7, 2014).
- Lieberman, R. C. and G. M. Shaw (2000). Looking inward, looking outward: The politics of state welfare innovation under devolution. *Political Research Quarterly* 53(2), 215–240.
- List, J. A. and D. M. Sturm (2006). How elections matter: Theory and evidence from environmental policy. *Quarterly Journal of Economics* 121(4), 1249–1281.
- Majumdar, S. and S. W. Mukand (2004). Policy gambles. *American Economic Review* 94(4), 1207–1222.
- Mead, L. M. (2004). *Government Matters. Welfare Reform in Wisconsin*. Princeton University Press.
- Meyer, B. D. and J. X. Sullivan (2008). Changes in the consumption, income, and well-being of single mother headed families. *American Economic Review* 98(5), pp. 2221–2241.
- Moen, E. R. and C. Riis (2010). Policy reversal. *American Economic Review* 100(3), 1261–68.
- Moffitt, R. A. (2001). The effect of pre-PRWORA waivers on AFDC caseloads and female earnings, income, and labor force behavior. In S. Danziger (Ed.), *Economic Conditions and Welfare Reform*. W.E. Upjohn Institute.
- Moffitt, R. A. (2002). Welfare programs and labor supply. In A. J. Auerbach and M. Feldstein (Eds.), *Handbook of Public Economics*, Chapter 34, pp. 2393–2430. Elsevier.
- Moffitt, R. A. (2008). Welfare reform: The US experience. *Swedish Economic Policy Review* 14 (2), 11–54.
- Oates (1999). An essay on fiscal federalism. *Journal of Economic Literature* 37(3), 1120–1149.
- Persson, T. and G. E. Tabellini (2000). *Political Economics: Explaining Economic Policy*. MIT Press.
- Reintsma, M. (2007). *The Political Economy of Welfare Reform in the United States*. Edward Elgar.
- Rogoff, K. (1990). Equilibrium political budget cycles. *American Economic Review* 80(1), 21–36.
- Rose-Ackerman, S. (1980). Risk taking and reelection: Does federalism promote innovation? *Journal of Legal Studies* 9, 593–616.
- Saavedra, L. (2000). A model of welfare competition with empirical evidence from AFDC. *Journal of Urban Economics* 47, 248–279.
- Smart, M. and D. M. Sturm (2013). Term limits and electoral accountability. *Journal of Public Economics* 107, 93–102.
- Soss, J., S. F. Schram, T. P. Vartanian, and E. O'Brien (2001). Setting the terms of relief: Explaining state policy choices in the devolution revolution. *American Journal of Political Science* 45(2), 378–395.
- Strumpf, K. S. (2002). Does government decentralization increase policy innovation? *Journal of Public Economic Theory* 4(2), 207 – 241.
- United States Census Bureau (2011). Statistical Abstract of the United States: 2012 (131st edition).

- Urban Institute (2015). Welfare Rules Database. Washington, D.C. (accessed on May 7, 2015).
- U.S. Department of Health and Human Services, Administration for Children and Families (1998-2009). TANF Annual Reports to Congress, I-VIII. Technical report, GAO.
- U.S. Department of Homeland Security (2011). *Yearbook of Immigration Statistics*. Washington, D.C.: Office of Immigration Statistics.
- Walker, J. L. (1969). The diffusion of innovations among the American States. *American Political Science Review* 63(3), 880–899.
- Weaver, K. (2000). *Ending Welfare as we Know it*. Brookings Institution.
- Weissert, C. S. (2000). *Learning from Leaders: Welfare Reform Politics and Policy in five Midwestern States*. Rockefeller Institute Press.
- Wheaton, W. C. (2000). Decentralized welfare: Will there be underprovision? *Journal of Urban Economics* 48, 536–555.
- Willems, T. (2013). Political accountability and policy experimentation: Why to elect left-handed politicians? *Mimeo. Princeton University*. mimeo.
- Winston, P. (2002). *Welfare Policymaking in the States: The Devil in Devolution*. Georgetown University Press.

Table 1: Initial Reputation, Lame Ducks and Policy Experimentation during the Waiver Period

	<u>Main Effect Reputation</u>		<u>Differential Effect for Lame Ducks</u>		<u>Control for Potential Gains</u>	
	(1)	(2)	(3)	(4)	(5)	(6)
Vote Margin Last Gubernatorial Election	-0.0071** (0.003)	-0.0076** (0.004)	-0.0087* (0.004)	-0.0093** (0.004)	-0.0084** (0.004)	-0.0089** (0.004)
Lame Duck * Gubernatorial Vote Margin			0.0049 (0.006)	0.0056 (0.007)	0.0062 (0.007)	0.0068 (0.007)
Lame Duck			0.0086 (0.076)	-0.0116 (0.082)	-0.0313 (0.099)	-0.0466 (0.101)
Past AFDC Spending					0.0004*** (0.000)	0.0004*** (0.000)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes
Switch of Governor Party (Past 2 Elections)	No	Yes	No	Yes	No	Yes
Observations	738	738	738	738	738	738
R-squared	0.292	0.294	0.293	0.295	0.311	0.312

Notes : The dependent variable in all specifications is the number of policy experiments (waiver applications) in a given state and year. Columns (1) and (2) include the governor's vote margin in the last gubernatorial election (measured as the distance to the runner-up in percent). Columns (3) and (4) allow the effect of the vote margin to differ for governors who cannot be reelected ("lame ducks"). Columns (5) and (6) also control for lagged AFDC spending (measured in US\$ million) as a proxy for the potential gains from experimentation. All specifications control for year and state fixed effects as well as governor age. Even columns also include an indicator whether the party of the governor switched (e.g. from Democrat to Republican or vice versa) over the past two electoral cycles to control for swing states. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table 2: Governor Experience and Policy Experimentation during the Waiver Period

	<u>Main Effect</u> <u>Governor's Experience</u>		<u>Differential Effect</u> <u>for Lame Ducks</u>		<u>Specification with</u> <u>all Controls</u>	
	(1)	(2)	(3)	(4)	(5)	(6)
Governor with Little Experience	-0.0413 (0.075)	-0.0368 (0.075)	-0.0929 (0.087)	-0.0900 (0.087)	-0.0974 (0.089)	-0.0893 (0.087)
Lame Duck * Governor with Little Experience			0.1977** (0.098)	0.1968** (0.098)	0.2113** (0.104)	0.2141** (0.103)
Lame Duck			-0.0376 (0.059)	-0.0203 (0.061)	-0.1175 (0.123)	-0.1344 (0.125)
Past AFDC Spending					0.0004*** (0.000)	0.0004*** (0.000)
Vote Margin Last Gubernatorial Election					-0.0084** (0.004)	-0.0089** (0.004)
Lame Duck * Gubernatorial Vote Margin					0.0066 (0.007)	0.0071 (0.007)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes
Switch of Governor Party (Past 2 Elections)	No	Yes	No	Yes	No	Yes
Observations	738	738	738	738	738	738
R-squared	0.288	0.292	0.292	0.297	0.315	0.317

Notes : The dependent variable in all specifications is the number of policy experiments (waiver applications) in a given state and year. Columns (1) and (2) include an indicator whether a governor is below the median political experience in the governor sample (as a proxy for the governor's knowledge). Columns (3) and (4) allow the effect of governor experience to differ for governors who cannot be reelected ("lame ducks"). Columns (5) and (6) shows the full specifications which also controls for lagged AFDC spending (measured in US\$ millions) as a proxy for the potential gains from experimentation; the governor's vote margin in the last gubernatorial election (measured as the distance to the runner-up in percent) and an indicator whether the governor cannot be reelected as a proxy for reelection concerns ("lame duck"). All specifications control for year and state fixed effects as well as governor age. Even columns also include an indicator whether the party of the governor switched (e.g. from Democrat to Republican or vice versa) over the past two electoral cycles to control for swing states. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table 3: Policy Experimentation during the TANF Period

	<u>Reputation Effect</u>			<u>Governor's Experience</u>	<u>Specification with all Controls</u>
	(1)	(2)	(3)	(4)	(5)
Vote Margin Last Gubernatorial Election	-0.0052** (0.002)	-0.0075** (0.003)	-0.0078** (0.003)		-0.0065** (0.003)
Lame Duck * Gubernatorial Vote Margin		0.0052 (0.005)	0.0056 (0.005)		0.0056 (0.005)
Lame Duck		-0.0414 (0.107)	-0.0510 (0.105)	-0.0020 (0.072)	-0.0577 (0.126)
Past TANF Spending			0.0003** (0.000)		0.0000 (0.000)
Governor with Little Experience				-0.0190 (0.075)	-0.0126 (0.091)
Lame Duck * Governor with Little Experience				0.0837 (0.124)	0.0890 (0.130)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes
Switch of Governor Party (Past 2 Elections)	Yes	Yes	Yes	Yes	Yes
Observations	544	544	544	544	544
R-squared	0.337	0.345	0.348	0.334	0.353

Notes : The dependent variable in all specifications is the number of policy experiments (changes in welfare rules) in a given state and year during the TANF period (1996-2010). Columns (1) and (2) include the governor's vote margin in the last gubernatorial election (measured as the distance to the runner-up in percent). Columns (3) and (4) allow the effect of the vote margin to differ for governors who cannot be reelected ("lame ducks"). Columns (5) and (6) also control for TANF spending (measured in US\$ million) two years earlier as a proxy for the potential gains from experimentation. All specifications control for year and state fixed effects as well as governor age. Even columns also include an indicator whether the party of the governor switched (e.g. from Democrat to Republican or vice versa) over the past two electoral cycles to control for swing states. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table 4: Electoral Incentives and Policy Reversals in the TANF Period

	<u>Main Effect</u> <u>Governor's Experience</u>		<u>Effect for</u> <u>Potential Gains</u>		<u>Specification with</u> <u>all Controls</u>	
	(1)	(2)	(3)	(4)	(5)	(6)
Governor with Little Experience	0.0504** (0.021)	0.0497** (0.021)	0.0576** (0.023)	0.0563** (0.023)	0.0569** (0.023)	0.0568** (0.024)
Governor Little Experience * Past TANF Spending			-0.0131** (0.005)	-0.0119** (0.005)	-0.0133** (0.005)	-0.0118** (0.006)
Past TANF Spending			0.0344 (0.037)	0.0306 (0.039)	0.0342 (0.037)	0.0306 (0.039)
Lame Duck					0.0038 (0.023)	-0.0028 (0.025)
Vote Margin Last Gubernatorial Election		0.0011 (0.001)		0.0011 (0.001)		0.0011 (0.001)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes
Observations	585	585	585	585	585	585
R-squared	0.116	0.119	0.117	0.120	0.117	0.120

Notes : The dependent variable in all specifications is an indicator whether the state has reversed a policy rule in a given year or not during the TANF period (1997-2010). Columns (1) and (2) include an indicator whether a governor has below median political experience in the governor sample to proxy for a governor's competence. Columns (3) and (4) include past state TANF spending (measured in US\$ million) as a measure of the potential gains from experimentation. Columns (5) and (6) also add an indicator whether the governor cannot be reelected ("lame duck"). Even columns also add the vote margin in the past gubernatorial election as a proxy for initial reputation (measured as the distance to the runner-up in percent). All specifications control for state and year fixed effects as well as governor age. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table 5: The Role of Ideology for Policy Experimentation

	(1)	(2)	(3)	(4)	(5)	(6)
Governor's Past Vote Margin	-0.0066*	-0.0065*	-0.0065*	-0.0065*	-0.0066*	-0.0066*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Lame Duck * Past Vote Margin	0.0086	0.0080	0.0086	0.0086	0.0085	0.0080
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Lame Duck	-0.1187	-0.0823	-0.1191	-0.1185	-0.1172	-0.0914
	(0.107)	(0.108)	(0.106)	(0.107)	(0.106)	(0.109)
Past TANF Spending	0.0003***	0.0003***	0.0003***	0.0003***	0.0003***	0.0003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Governor with Little Experience	-0.0973	-0.1010	-0.0987	-0.0977	-0.0971	-0.0897
	(0.081)	(0.084)	(0.082)	(0.081)	(0.079)	(0.068)
Lame Duck * Governor with Little Experience	0.2017**	0.2055**	0.2024**	0.2023**	0.2005**	0.2078**
	(0.095)	(0.096)	(0.095)	(0.094)	(0.093)	(0.100)
Republican Governor	0.0035	0.0142				-0.0534
	(0.039)	(0.042)				(0.159)
Lameduck * Republican Governor		-0.0700				-0.0643
		(0.072)				(0.076)
Governor Party Ideology (Berry et al. 1998)			0.0077			0.1249
			(0.082)			(0.310)
Citizens Ideology (Berry et al. 1998)				-0.0003		-0.0000
				(0.003)		(0.004)
Democratic Presidential Vote Share					0.0015	0.0011
					(0.005)	(0.005)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes
Observations	785	774	785	785	785	774
R-squared	0.296	0.298	0.296	0.296	0.297	0.298

Notes : The dependent variable in all specifications is the number of policy experiments (waiver period) in a given state and year. All main variables are defined as before (see notes to Tables 1 and 2 for details). The ideology measure for the government is calculated from Berry et al. (1998) and ranges from zero for most conservative to 100 for most liberal. The citizen ideology measure is also taken from Berry et al. (1998); see the data appendix for further details. Democratic presidential vote share refers to the last presidential election. All specifications control for year and state fixed effects as well as governor age. Standard errors clustered at the state level are shown in parentheses.

*** p<0.01, ** p<0.05 and *p<0.1.

Table 6: Composition of Legislature and Policy Experimentation

	(1)	(2)	(3)	(4)	(5)
Governor's Past Vote Margin	-0.0059* (0.004)	-0.0064** (0.004)	-0.0058 (0.004)	-0.0061* (0.004)	-0.0055 (0.004)
Lame Duck * Past Vote Margin	0.0086 (0.007)	0.0091 (0.007)	0.0099 (0.007)	0.0088 (0.007)	0.0097 (0.007)
Lame Duck	-0.1141 (0.106)	-0.1165 (0.107)	-0.1294 (0.112)	-0.1198 (0.109)	-0.1218 (0.112)
Past TANF Spending	0.0003*** (0.000)	0.0003*** (0.000)	0.0003*** (0.000)	0.0003*** (0.000)	0.0003*** (0.000)
Governor with Little Experience	-0.1122 (0.081)	-0.0999 (0.081)	-0.0963 (0.083)	-0.0984 (0.082)	-0.1084 (0.085)
Lame Duck * Governor with Little Experience	0.1952** (0.096)	0.1989** (0.095)	0.2041** (0.098)	0.2021** (0.095)	0.2001* (0.100)
Dem. Seat Share Upper House	-0.3099 (0.306)				-0.1606 (0.293)
Dem. Seat Share Lower House	-0.0529 (0.352)				-0.0156 (0.344)
Divided Government		0.0657 (0.044)			0.0411 (0.046)
Polarization Upper House			-0.5426 (0.377)		-0.4702 (0.410)
Polarization Lower House			-0.1779 (0.386)		-0.1348 (0.382)
Competitiveness Index				0.0045 (0.004)	-0.0015 (0.005)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes
Observations	785	778	785	776	769
R-squared	0.298	0.299	0.300	0.296	0.301

Notes : The dependent variable in all specifications is the number of policy experiments (waiver period) in a given state and year. All main variables are defined as before (see notes to Tables 1 and 2 for details). Divided Government is equal to one if the party of the governor is different from the party of the majority of legislators in either the state's lower or upper house. The polarization variables are calculated as absolute deviations of the democratic seat share from 50%. Competitiveness (based on Holbrook and Van Dunk (1993)) is measured as the average voter support in percent that winning candidates in state legislative elections obtained over the current and past 3 years. All specifications control for state and year fixed effects as well as governor age. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table 7: Spillovers between States and Policy Experimentation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Governor's Past Vote Margin	-0.0082** (0.004)	-0.0083** (0.004)	-0.0084** (0.004)	-0.0084** (0.004)	-0.0085** (0.004)	-0.0059 (0.005)	-0.0080* (0.004)
Lame Duck * Past Vote Margin	0.0063 (0.007)	0.0078 (0.007)	0.0061 (0.007)	0.0076 (0.007)	0.0067 (0.007)	0.0074 (0.009)	0.0068 (0.007)
Lame Duck	-0.1157 (0.123)	-0.1441 (0.127)	-0.1165 (0.124)	-0.1407 (0.127)	-0.1213 (0.122)	-0.1999 (0.139)	-0.1368 (0.130)
Past TANF Spending	0.0004*** (0.000)	0.0005*** (0.000)	0.0004*** (0.000)	0.0004*** (0.000)	0.0004*** (0.000)	0.0007*** (0.000)	0.0004*** (0.000)
Governor with Little Experience	-0.0986 (0.089)	-0.1377 (0.096)	-0.0993 (0.086)	-0.1346 (0.096)	-0.0963 (0.089)	-0.0694 (0.111)	-0.1398 (0.094)
Lame Duck * Governor with Little Experience	0.2148** (0.103)	0.2147* (0.112)	0.2190** (0.103)	0.2095* (0.113)	0.2141** (0.103)	0.1245 (0.128)	0.2284** (0.109)
Geographic Neighbors' Experiments (t-1)	-0.1247 (0.088)						-0.1650* (0.098)
Geographic Neighbors' Experiments (t-2)		-0.0898 (0.092)					-0.1196 (0.096)
Pop. Size Neighbors' Experiments (t-1)			0.1753 (0.207)				0.1759 (0.209)
Pop. Size Neighbors' Experiments (t-2)				0.0388 (0.169)			0.0485 (0.161)
Ideological Neighbors' Experiments (t-1)					-0.0514 (0.101)		0.0063 (0.090)
Ideological Neighbors' Experiments (t-2)						0.1924*** (0.072)	-0.0027 (0.072)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	738	691	738	691	738	691	691
R-squared	0.318	0.319	0.320	0.319	0.316	0.193	0.328

Notes : The dependent variable in all specifications is the number of policy experiments (waiver period) in a given state and year. All main variables are defined as before (see notes to Tables 1 and 2 for details). Experiments by geographic neighbors refers to the average number of waivers for adjacent states (sharing a boundary). Experiments by population size neighbors refers to the average number of waivers in states with similar population size (where states are divided into ten groups of similar population size). Experiments by ideological neighbors refers to the average number of waivers for the four states within the same census region which voted most similarly in the 1978 presidential election. All specifications control for year and state fixed effects as well as governor age. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table 8: The Demand for Welfare and Policy Experimentation

	(1)	(2)	(3)	(4)	(5)
Governor's Past Vote Margin	-0.0062** (0.004)	-0.0064** (0.004)	-0.0054 (0.004)	-0.0066** (0.004)	-0.0052 (0.004)
Lame Duck * Past Vote Margin	0.0077 (0.007)	0.0089 (0.007)	0.0082 (0.007)	0.0092 (0.007)	0.0086 (0.007)
Lame Duck	-0.1111 (0.106)	-0.1206 (0.108)	-0.1461 (0.114)	-0.1320 (0.107)	-0.1562 (0.118)
Past TANF Spending	0.0003 (0.000)	0.0003*** (0.000)	0.0003*** (0.000)	0.0003*** (0.000)	0.0003** (0.000)
Governor with Little Experience	-0.0935 (0.082)	-0.0971 (0.080)	-0.0903 (0.081)	-0.1005 (0.079)	-0.0869 (0.081)
Lame Duck * Governor with Little Experience	0.2070** (0.098)	0.2033** (0.095)	0.2283** (0.099)	0.2015** (0.093)	0.2295** (0.101)
Unemployment Rate	-0.0153 (0.014)				-0.0121 (0.016)
Per Capita Income (in \$1,000)	-0.0000 (0.000)				0.0000 (0.000)
Population (in 1,000)	-0.0211 (0.096)				0.0130 (0.084)
% Population Black		-0.0269 (0.066)			-0.0151 (0.066)
% Population 65+		-0.0142 (0.064)			-0.0156 (0.073)
% Immigrant Population			-0.0557*** (0.007)		-0.0568*** (0.011)
% Unmarried Births			-0.0095 (0.012)		-0.0090 (0.013)
Income Inequality				-0.0423** (0.020)	-0.0369* (0.021)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes
Observations	785	785	785	785	785
R-squared	0.297	0.297	0.305	0.299	0.308

Notes : The dependent variable in all specifications is the number of policy experiments (waiver period) in a given state and year. All main variables are defined as before (see notes to Tables 1 and 2 for details). Income inequality is measured as the 90th percentile over the 10th percentile within each state and year. All specifications control for state and year fixed effects as well as governor age. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Figure 1: Evolution of Welfare Waivers, 1978-1996

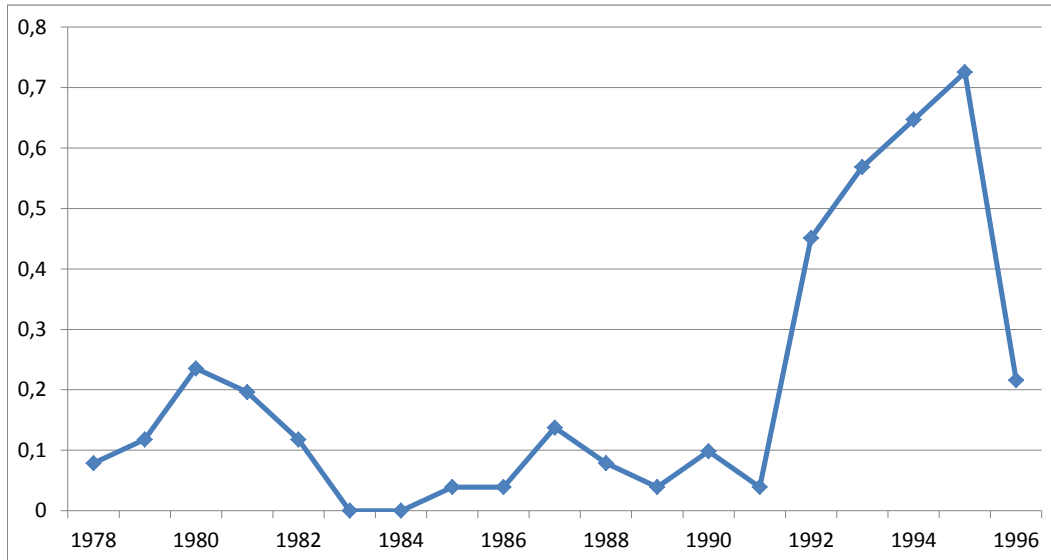


Figure 2: Distribution of Waiver Applications across States, 1978-1996

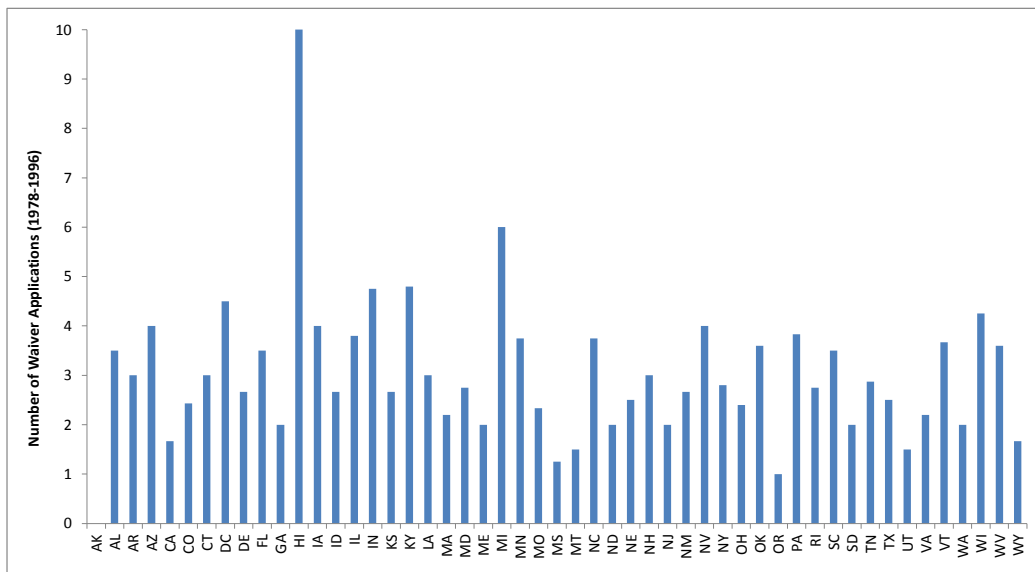
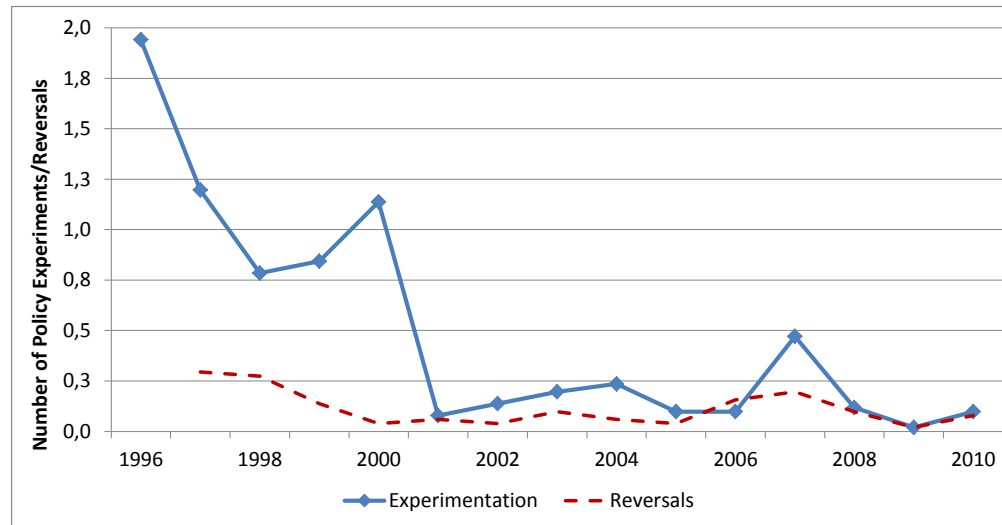
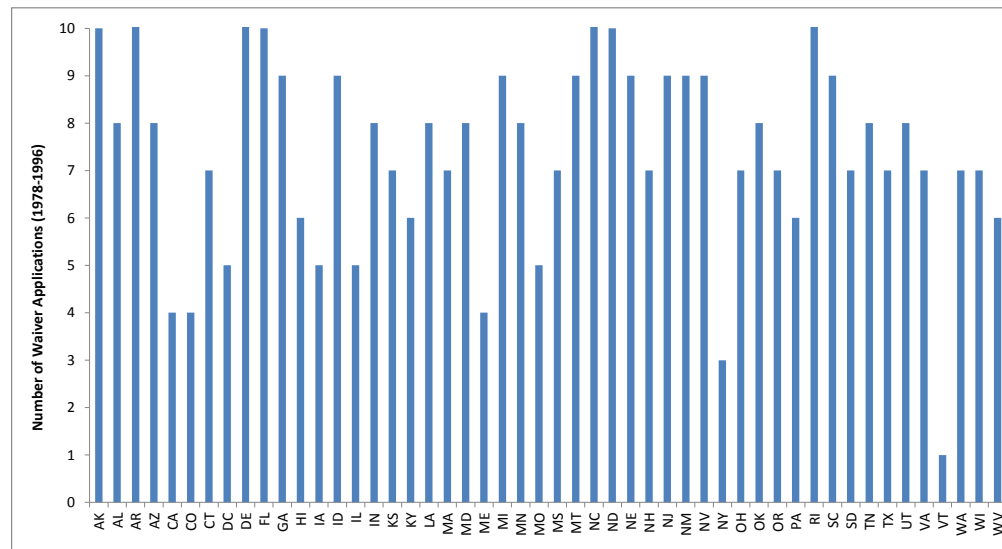


Figure 3: Evolution of Policy Experiments and Reversals, 1996-2010



Source : Authors' calculation; see the data appendix and Table A1 for a description of the welfare rules.

Figure 4: Distribution of Policy Experiments across States, 1996-2010



Source : Authors' calculation; see the data appendix and Table A1 for a description of the welfare rules.

Online Appendix (for Online Publication Only)

A Theory

A.1 Proof of Propositions

Proof of Proposition 1. The proof of Proposition 1 follows from the optimization problem of the incumbent politician. A high-quality incumbent politician knows the state of the world. At $T = 0$, the incumbent chooses to reform or not (with the option to revert back to a_S if the reform is undertaken) in order to maximize equation (1).

In state S , no gain can be obtained from the reform ex-post: The ex-post welfare of the reform in state S is either $-c$ or $-\frac{c}{2}$ if reverted back to a_S . Therefore, the reform should not be undertaken from a welfare perspective. Since the ex-post welfare of a reform in state S is negative, the probability of reelection (from the reelection rule) in state S is (weakly) positive if and only if no reform is undertaken. The optimal decision of the high-quality incumbent is then to stick with the status quo policy a_S independently of γ and λ .

In state N , undertaking the reform maximizes ex-post welfare since there are positive welfare gains ex-post with probability 1, i.e. the ex-post welfare with a reform in state N is equal to $\Delta - c$ which is strictly positive from (A1), and is never reversed independently of the realization of the benefit in the short term. In addition, the probability of reelection is also maximized by implementing the reform since a positive ex-post welfare ensures reelection with probability 1, which is (weakly) greater than λ for any $\lambda \in [0, 1]$.

Proof of Proposition 2. The proof of Proposition 2 follows from the optimization problem of the incumbent politician given by equation (1) for $\gamma = 1$. For a low-quality politician who does not know the state of the world, the reform is welfare maximizing from an ex-ante perspective from (A1) and (A2). Therefore, she always implements the reform in $T = 0$. However, it is welfare-maximizing to revert back to a_S if the reform fails in the short run since the expected future benefit from the reform is negative from (A3).

Proof of Proposition 3. The proof of Proposition 3 follows from the optimization problem of the incumbent politician given by equation (1) for $\gamma = 0$.

A purely office-motivated politician never reverts back to the status quo after implementing a reform. If the reform delivers benefits in the short term it is always continued. If the reform does not deliver short-term benefits in $T = 1$, a policy reversal back to the status quo implies that the politician would lose the election with probability 1 (because welfare ex-post would be equal to $-\frac{c}{2}$ and hence, the probability of reelection would be zero from the reelection rule). Whereas if the politician persists with the reform, she can win the election for sure if the long-term benefit realizes which happens with probability $\beta > 0$.

For a low-quality politician who does not reverse a reform, the reform maximizes the probability of winning from an ex-ante perspective if it delivers higher chances of reelection than her current reputation. This statement is equivalent to equation (2) where the left-hand side represents the ex-ante winning probability if the reform is undertaken and continued which is equivalent to the probability that the state of the world is N . The

right-hand side is the ex-ante winning probability if the reform is not undertaken which is equal to the initial reputation of the incumbent.

Proof of Corollary 1. The proof of Corollary 1 follows from equation (2).

Proof of Proposition 4. The proof of Proposition 4 directly follows from comparing the politician's payoffs in case of reverting the reform and of continuing the reform after a short-term failure. Reverting the reform is at least as good as continuing it if and only if

$$\gamma \left(\frac{-c}{2} \right) \geq \gamma (\beta \Delta - c) + (1 - \gamma) \beta,$$

or

$$-\beta \geq \gamma \left(\beta \Delta - \frac{c}{2} - \beta \right).$$

Since the term in parenthesis is negative from (A3), the last inequality can be rewritten to obtain the threshold γ^* in equation (3).

Proof Corollary 2. The proof of Corollary 2 directly follows from differentiating equation (3) with respect to γ and Δ , respectively.

A.2 Intermediate politician and decision to experiment

Here, we analyze the case whether a low-quality incumbent who cares about both welfare and reelection ($\gamma \in [0, 1]$) undertakes the reform.

Proposition 5. (I.) If $\gamma \geq \gamma^*$, then an incumbent politician with low-quality always undertakes a reform if and only if

$$\gamma \left(pq\Delta - \frac{c}{2}(1 + pq) \right) + (1 - \gamma)pq \geq (1 - \gamma)\lambda. \quad (7)$$

(II.) If $\gamma < \gamma^*$, then an incumbent politician with low-quality always undertakes a reform if and only if

$$\gamma(p\Delta - c) + (1 - \gamma)p \geq (1 - \gamma)\lambda. \quad (8)$$

Proof. The proof of Proposition 5 directly follows from the comparison of the politician's payoff if she undertakes the reform (for the two cases where she will or will not revert back to the status quo in the face of short-term failure) with the payoff from sticking with the status quo and running the next election on her initial reputation instead. $\square$

Proposition 5 shows that a low-quality incumbent compares the possible welfare and reelection gains that experimentation will deliver (taking into account her optimal reversal decision) to her current chances of reelection determined by her initial reputation.

Corollary 3. A low-quality incumbent is more likely to experiment when Δ increases, γ increases and λ decreases.

Proof. The proof of Corollary 3 follows directly from differentiating equations (7) and (8) with respect to Δ , γ and λ respectively. $\square$

B Data

B.1 Welfare Policy Rules during the TANF period

To measure policy experimentation and reversals for the post-1996 period, we rely on the Welfare Rules Database by the Urban Institute. The Welfare Rules Database (WRD) was developed to provide detailed information about states' TANF policies obtained from caseworker manuals and regulations. The database is often more detailed and up-to-date than the official plans of state choices under the block grant that states submit periodically to the federal government. The database contains literally hundreds of rules on eligibility, benefit calculation, and many other aspects of welfare reform. We focus on a set of rules in policy areas that were at the center of the public and political debate surrounding welfare reform. Table A1 provides an overview of the rules and how we code policy experiments and policy reversals.

Family caps. Under AFDC, benefit levels increased with family size. Hence, when a child was born to a member of an assistance unit, the benefit increased to meet the needs of the new child. Under a family cap, additional benefits an assistance unit would receive for adding another member to the unit would be capped. Some states provide a percentage of the increase to the unit, while others provide no additional funds to the unit for the additional child. In several states, a family is never able to regain benefits for a capped child, even after the case has been closed for a period of time. In others, a family cap can be removed (and hence, the child can be included in the benefit computation should the family apply for assistance again) if the assistance unit remains off welfare for some time. We code a binary indicator equal to one if a state has a family cap and zero if not.

Work requirements. Under AFDC, states could require recipients to participate in the Job Opportunities and Basic Skills Training (JOBS) program, which provided education, training, and work experience activities. However, many individuals were exempt from these requirements (because of age, illness or having a small child). Under TANF, states require adults heading an assistance unit to perform some type of work-related activity. Work programs vary widely from state to state in terms of who must work, how much work is required, and what activities are considered work. The first rule defines the minimum number of hours a recipient must participate in work-related activities. The hours requirements vary from a mere effort to find a job up to full-time employment. The second rule defines whether the work requirement applies after several months of benefit receipt or by the time of application or approval. The third rule indicates whether there is a time limit of benefit receipt if a parent fails to work at least 20 hours per week in a regular job after a certain number of assistance months. The fourth rule counts the number of work exemptions due to, for example, pregnancy, disability or caring for a young child or elderly person.

Sanctions. Under AFDC, sanctions were sometimes imposed if the adult heading an assistance unit did not comply with the JOBS program; in that case, the adult portion of the benefit was not paid out to the assistance unit. Under TANF, states now require household heads to perform some type of work-related activity sooner or later. If a benefit unit does not comply with these requirements, states can impose drastic sanctions. The

first policy rule characterizes the initial sanction if a benefit unit fails to comply with the work requirements for the first time. The initial sanction varies from reduction of 25% or less to a suspension of the full family benefit. The other three rules characterize the severity of the worst sanction that can be imposed. The second rule defines the severity of the worst sanction varying from less than 25% of the benefit to a suspension of the full family benefit and even case closure. The third rule defines the duration of the worst sanction ranging from until the unit complies with the requirement to a permanent suspension of the family benefit. The final rule defines whether a unit has to reapply (or not) after the worst sanction has been imposed.

Termination and work-requirement time limits. Under AFDC, families were entitled to receive benefits as long as they met the eligibility requirements. Under TANF, many states imposed both intermittent and lifetime time limits. The first rule characterizes the number of months an assistance unit can receive benefits without interruption (many states impose 24 months). The second rule defines how benefits are reduced when the assistance unit reaches the intermittent time limit. The loss of benefits might be just for the adult members or for the entire assistance unit. The third rule defines the duration of the lifetime limit ranging from no time limit to only 24 months. The federal government has itself imposed a 60 months lifetime limit; states that wish to extend benefit receipt beyond five years have to use their own state funds to finance it. The final rule defines whether the state allows for any extensions to the lifetime limit or not.

B.2 Politics and ideology measures

Governor characteristics. Information on U.S. governors comes from the website of the National Governor’s Association merged with data provided by David J. Andersen from the Eagleton Institute of Politics at Rutgers University. We calculate a governor’s competence from her political experience prior to becoming governor. Specifically, the variable measures the number of years between the first electoral office (such as member of the State Senate or Attorney General) and the inauguration as governor. We then construct a dummy variable equal to one if a governor has prior political experience below the median and zero otherwise. We further code information on the age of the governor from the same sources. Information on gubernatorial election results (especially vote margins) and term limits are taken from List and Sturm (2006); Council of State Governments (2012); and Leip (2012).

State Legislature. Data for the composition of the state legislature, state competitiveness, the party of the governor and indicators for a divided government are obtained from Klarner (2003) and updated using Klarner’s webpage (<http://www.indstate.edu/polsci/klarnerpolitics.htm>). Polarization in the state legislature is calculated as $|\text{democratic seat share} - 50\%|$ for the state senate and house respectively. The divided government indicator is equal to one if the governor belongs to a different party than the majority of legislators in either the state senate or the state house. To measure state competitiveness, we use the Holbrook-Van Dunk index calculated from district-level returns to state legislative elections (see Holbrook and Van Dunk (1993) for details).

Ideology. The vote share for the Democratic candidate in the last presidential election is taken from the Statistical Abstract (United States Census Bureau (2011)) and updated using Leip (2012). We also use voter and government ideology based on ideology ratings of the state’s congressional delegation, the American for Democratic Action (ADA) rating and the AFL/CIO’s Committee on Political Education (COPE) rating (Berry et al. (1998)). Berry et al. (1998) assign an ideology rating to the citizens of each congressional district using a weighted average of the score of the congressional member and his or her election opponent, weighting the scores according to the number of votes they received. Zero denotes the most conservative and 100 the most liberal. They then generate a state-wide measure by averaging over all congressional districts. The measure of government ideology is constructed by assigning to the governor and major party delegations in the legislature the ratings of the members of Congress from their party. Updates of these ideology data are available at <http://rcfording.wordpress.com/state-ideology-data/>.

B.3 State Demographics and Other Controls

Demographics. Population size, the number of Blacks and the age structure are taken from the Statistical Abstract (United States Census Bureau (2011)). The size of the immigrant population refers to the number of legal immigrants admitted by state of intended residence and is taken from Fang and Keane (2004) for 1970 to 2002 and updated using the Yearbook of Immigration Statistics (U.S. Department of Homeland Security (2011)) which is available online at <http://www.dhs.gov/yearbook-immigration-statistics>. Personal income per capita is taken from the website of the Bureau of Economic Analysis. The unemployment rate for 1960 to 1998 is from the website of the Bureau of Labor Statistics. Our measure of income inequality is the ratio of the 90th over the 10th percentile of total household income calculated from the March Current Population Survey (Center for Economic and Policy Research (2012)). All income variables are deflated by the urban consumer price index with base year 2002. Unmarried Birth refers to the % of all births to unmarried women per 1,000 unmarried women aged 15-44 years by state of residence. For the years 1992 to 2003, the data are available from Table 8.3 in the TANF Annual Reports to Congress (U.S. Department of Health and Human Services, Administration for Children and Families (2009)). For earlier and later years, we obtain the data from the Center for Disease Control and Prevention (CDC), the National Vital Statistics System and Vital Statistics available from the National Bureau of Economic Research.

State AFDC and TANF spending. To measure the potential gains from experimentation, we use state-level AFDC and TANF spending. These data come from Paul Ehmann at the U.S. Census Bureau. All fiscal variables are then converted into real measures using the urban consumer price index (with years 1982-1984=100) from the Bureau of Labor Statistics.

Spillover effects across States. To analyze spillover effects across states, we define a “neighbor” along three different dimensions. The first variable (geographic neighbors) codes the extent of experimentation for all states sharing a common border with the current state. The second variable (similar population sizes) calculates experimentation in states with a population size similar to that of the current state. For the calculation, we use the following ten bands based on population size in 1978: (CA NY TX PA IL),

(OH MI FL NJ MA), (NC IN GA VA MO), (WI TN MD LA MN), (WA AL KY CT SC), (IA OK CO AZ OR), (MS KS AR WV NE), (UT NM ME RI HI), (ID NH MT NV SD) and (ND DE VT WY AK). The third measure (ideological neighbors) takes the average number of experiments of the four states within the same census region who voted similarly in the presidential election of 1978.

Table A1: Coding of Policy Experimentation and Reversals during the TANF Period (1996-2010)

Policy Rule	Description of Rule	Experimentation = 1	Reversal = 1
<u>Family Cap (1)</u>	Benefits do not increase if an additional child is born in family while receiving benefits.	State adopts family cap before or after 1996.	State abolishes family cap in 1997 or later.
<u>Work Requirements (3)</u>			
Hours Requirement	Minimum # of hours a recipient must participate in work-related activities.	Increase in minimum number of working hours required of the recipient.	State reduces or abolishes minimum number of working hours required of the recipient.
Work upon Enrollment	Work requirements apply at application, approval or upon benefit receipt (or only later).	State requires work requirement to be fulfilled at application, approval or upon benefit receipt.	State allows work requirement to be fulfilled later.
Time Limit to Work	Work at least 20 hours per week in an unsubsidized job after a certain period of benefit receipt.	State adopts a time limit to fulfill the requirement of working in a unsubsidized job for at least 20 hours.	State abolishes time limit to fulfill the requirement of working in an unsubsidized job for at least 20 hours.
<u>Time Limits (3)</u>			
Duration of Lifetime Limit	Maximum # of months an assistance unit can receive benefits over the lifetime.	Prior to 1997, if a time limit is adopted; after 1997, when a state adopts a time limit stricter than the 60 months federal time limit. In 1997, if a state adopts a time limit other than the federal time limit.	If a state abolishes the time limit or increases # of months.
Intermittent Time Limit	# months time an assistance units can receive benefits without interruption.	State reduces intermittent time limit.	State increases intermittent time limit.
Benefit Reduction after Intermittent Time Limit	How much benefits are reduced (adult portion or benefit unit) when assistance units hits the	State reduces benefits more after unit hits intermittent time limit.	State reduces benefits less after unit hits intermittent time limit.
<u>Sanctions (4)</u>			
Severity of Worst Sanction	How much benefits are reduced under worst sanction for non-compliance with work	State imposes a more severe penalty for non-compliance.	State reduces severity of worst sanction for non-compliance.
Duration of Worst Sanction	Length of most severe sanction for not complying with work requirements.	State increases duration of worst sanction.	State decreases duration of worst sanction.
Reapply	Whether the unit has to reapply after worst sanction for non-compliance has been imposed.	State adopts requirement that unit has to reapply after worst sanction has been imposed.	State abolishes requirement that unit has to reapply after worst sanction has been imposed.
Severity of Initial Sanction	Whether initial sanction for non-compliance is removal of full family benefit.	State adopts that full family benefit is removed as initial sanction.	State abolishes rule that full family benefit is removed as initial sanction.

Source : Welfare Rules Database (2015)

Table A2: Summary Statistics

	Waiver Period		Tanf Period	
	Mean	Std. Dev.	Mean	Std. Dev.
# Policy Experiments (Waiver Applications)	0,201	0,577		
Dummy for Policy Experiment (Waiver Application)	0,150	0,357		
# Policy Experiments (TANF period)			0,497	0,908
Dummy for Policy Reversal (TANF Period)			0,112	0,351
Governor's Past Vote Margin	8,497	7,316	11,427	10,406
Governor Lane Duck	0,238	0,426	0,303	0,460
Governor with Little Experience	0,491	0,500	0,462	0,499
Governor Age	52,840	7,856	55,616	7,576
Governor Party	0,587	0,489	0,063	0,318
Governor Party Ideology (Berry et al. 1998)	-0,019	0,279	0,063	0,318
Citizens Ideology (Berry et al. 1998)	47,400	15,210	51,282	15,989
Democratic Presidential Vote Share	42,725	9,050	47,620	9,925
Divided Government	0,555	0,497	0,547	0,498
Democratic Seat Share Upper House	0,598	0,185	0,512	0,162
Democratic Seat Share Lower House	0,589	0,179	0,519	0,154
Polarization Upper House	0,168	0,125	0,131	0,096
Polarization Lower House	0,162	0,118	0,125	0,091
Competitiveness Index (Holbrook and Van Dunk)	40,929	12,200	38,991	10,978
AFDC Spending (in US\$ millions)	330,4	683,9		
AFDC Spending (per Capita)	52,993	33,447		
TANF Spending (in US\$ millions)			490,430	990,560
TANF Spending (per Capita)			73,268	49,248
Unemployment Rate	6,466	2,110	5,280	1,848
Per Capita Income (/1000)	15719,2	5310,9	32342,4	7619,6
Population (/1000)	4,782	5,199	5,686	6,326
% Black Population	9,399	9,157	10,772	10,619
% Population 65+	11,941	2,272	12,722	1,805
% Immigrant Population	1,762	2,216	2,263	1,739
% Unmarried Births	24,096	8,812	35,518	7,428
Income Inequality (P90/P10)	7,895	1,741	8,833	1,992
Geographic Neighbors' Experiments (t-1)	0,220	0,409	0,341	0,452
Pop. Size Neighbors' Experiments (t-1)	0,203	0,351	0,501	0,658
Ideological Neighbors' Experiments (t-1)	0,229	0,393	0,492	0,660

Notes : For details on the welfare experimentation and reversal measures, see the data appendix. The quality of a governor is measured by years of political experience prior to becoming governor. The low quality governor indicator is equal to one if the number of years of political experience is below the median value. Governor Lane duck is equal to one if the governor cannot run for reelection. The past vote margin measures the winner's distance in votes to the runner-up in the last gubernatorial election. The ideology measure for the government is calculated from Berry et al. 1998 and ranges from zero for most conservative to 100 for most liberal, see the data appendix for details. The citizen ideology measure is also taken from Berry et al. 1998. Democratic Presidential Vote Share refers to the share in the last presidential election. Divided Government is equal to one if the party of the governor is different from the party of the majority of legislators in either the state's lower or upper house. The polarization variables are calculated as absolute deviations of the democratic seat share from 50%. The competitiveness index is the vote share of winning candidates in state legislative elections over the current and past 3 years. Geographic Neighbors' Experiments refers to the average number of TANF experiments for adjacent states (geographically neighboring states). Pop. Size Neighbors' Experiments refers to the average number of TANF experiments for states of similar population size (where all states are divided into ten bands of similar population size). Ideological Neighbors' Experiments refers to the average number of TANF experiments for the four states within the same census region which are ideologically most similar (in terms of 1978 presidential election results).

Table A3: Linear Probability Model of Policy Experimentation during the Waiver Period

Y: Applied for Waiver (Yes=1)	Reputation Effect			Governor's Political Experience		Specification with
	(1)	(2)	(3)	(4)	(5)	all Controls (6)
Past Governor Vote Margin	-0.0052** (0.002)	-0.0056** (0.002)	-0.0056** (0.002)			-0.0057** (0.002)
Lame Duck * Past Vote Margin		0.0015 (0.004)	0.0016 (0.004)			0.0019 (0.004)
Lame Duck		0.0058 (0.055)	0.0038 (0.057)		-0.0549 (0.037)	-0.0722 (0.069)
Past AFDC Spending			0.0000 (0.000)			0.0000 (0.000)
Governor with Little Experience				-0.0129 (0.036)	-0.0551 (0.039)	-0.0485 (0.038)
Lame Duck * Governor with Little Experience					0.1725*** (0.061)	0.1753*** (0.061)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes
Change in Governor Party (Past 2 Elections)	Yes	Yes	Yes	Yes	Yes	Yes
Observations	738	738	738	738	738	738
R-squared	0.321	0.322	0.322	0.314	0.322	0.330

Notes : The dependent variable in all specifications is equal to one if a state has applied for a welfare waiver in a given year and zero otherwise. All specifications report estimates from a linear probability model. The specifications in columns (1)-(3) correspond to columns (2), (4) and (6) in Table 1; the specifications in columns (4)-(6) correspond to columns (2), (4) and (6) in Table 2. All specifications control for year and state fixed effects, governor age and an indicator whether the party of the governor has changed (from Democrat to Republican or vice versa) over the past two electoral cycles. Standard errors clustered at the state level are shown in parentheses. *** p<0.01, ** p<0.05 and *p<0.1.

Table A4: Alternative Measures of Experimentation in the TANF Period

	<u>Restrict to First Change in Welfare Rule</u>				<u>Use Broader Set of Welfare Rules</u>			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Vote Margin Last Gubernatorial Election	-0.0036** (0.001)	-0.0069*** (0.002)	-0.0070*** (0.002)	-0.0068*** (0.002)	-0.0036 (0.003)	-0.0032 (0.004)	-0.0031 (0.004)	-0.0046 (0.007)
Lame Duck * Gubernatorial Vote Margin		0.0079 (0.005)	0.0080 (0.005)	0.0086 (0.005)		-0.0033 (0.007)	-0.0027 (0.007)	-0.0018 (0.004)
Lame Duck		-0.0888 (0.086)	-0.0904 (0.086)	-0.1177 (0.119)		0.1179 (0.143)	0.1022 (0.140)	0.1351 (0.167)
Past TANF Spending			0.0001 (0.000)	0.0003 (0.000)			0.0004* (0.000)	0.0001 (0.000)
Governor with Little Experience				-0.1187 (0.071)				-0.0217 (0.158)
Lame Duck * Governor with Little Experience				0.0967 (0.108)				0.0886 (0.093)
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Governor Age	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	544	544	544	492	544	544	544	492
R Squared	0.226	0.231	0.231	0.244	0.334	0.335	0.344	0.346

Notes : The table reports the same specification as columns (1)-(3) and (5) in Table 3 using alternative measures of policy experimentation during the TANF period. The first measure in columns (1)-(4) uses the same set of rules, but focuses on the first change in each welfare rule. The second alternative measure in columns (5)-(8) uses a broader set of welfare rules (including earnings disregards, mandatory job search and diversion payments) to define a policy experiment. See the data appendix and Table A1 for details on the coding of policy rules. See notes to Table 3 for details on the specification. Standard errors clustered at the state level are reported in parentheses. *** p<0.01, ** p<0.05 and * p<0.1.

Figure A1: Caseload per Capita, 1980-2010

