Are Judges Political Animals After All? Quasi-experimental Evidence from the German Federal Constitutional Court*

Thomas Gschwend†
University of Mannheim

Sebastian Sternberg‡
University of Mannheim

Steffen Zittlau§
University of Mannheim

Paper submitted for presentation at the first Conference on Empirical Legal Studies in Europe (CELSE), Amsterdam 2016

Abstract

Constitutional court judges maintain to be independent, apolitical actors, even though they get nominated by political elites. So far, much of the research has focused on the legal output of courts in order to show that judges are political animals. Studying outcomes is often plagued by endogeneity issues (e.g., when using votes to predict votes), or is simply not possible in some contexts due to non-disclosure of individual voting records. Alternatively, scholars employ a party-label heuristic and infer from the ideological position of the nominating party to the ideological position of judges. But do those ideological differences between judges become behaviorally relevant? In this paper, we provide two pieces of evidence that judges nominated by different parties seem to behave differently. First, we study the composition of three-judge panels (Chambers) of the German Federal Constitutional Court using Chamber decisions from 1998-2011 and show that homogenous panels have been actively avoided. Second, we analyze the court's replacement rules for absentee judges for the composition of chambers. We show that in situations where mechanically following the rules of procedure to replace absentee judges would lead to homogenous panels, the likelihood to deviate from those rules is systematically higher. Hence, judges elide formal rules to avoid homogenous chambers in order to appear politically unbiased.

*We like to thank Benjamin Engst, Erik Goetze, Wolfgang Rohrhuber, Daniel Stegmüller and Caroline Wittig for helpful discussions.

†Professor, Department of Political Science, University of Mannheim, A 5, 6, D-68131, Mannheim, Germany (gschwend@uni-mannheim.de).

‡Graduat Student, Department of Political Science, University of Mannheim, A 5, 6, D-68131, Mannheim, Germany (sebastian.sternberg@gess.uni-mannheim.de).

§PhD Candidate, Department of Political Science, University of Mannheim, A 5, 6, D-68131, Mannheim, Germany (zittlau@uni-mannheim.de).

1 Introduction

Can constitutional court judges said to be "political" in the sense that their political convictions might influence how they make decisions? While this question can be answered affirmative for Supreme Court judges in the US, it is still a fundamental question in the larger field of comparative judicial politics (Dyevre, 2010). Many scholars and law professionals themselves would argue that constitutional court judges are outsiders to the political world and will follow the law and nothing but the law. Therefore, one is left wondering what those judges do if law cannot give much or seemingly conflicting guidance? Others would concede that their might be an ideological dimension in judicial decision-making. However, as long as the nomination and election procedure as well as the way votes are reported is less focused on the individual judges, the relationship between judges' convictions, i.e., their ideological positions, and their decisions is considerably weaker than what we observe in the US. This is particularly the case because other factors such as collegial pressures (Sunstein et al., 2006), legal doctrine (Bailey and Maltzman, 2011), public opinion as well as preferences of other political actors (Hönnige, 2009; Vanberg, 2005) might render individually held political views largely irrelevant for judicial decision-making.

While generally the cards seem to be stacked against answering the above question affirmative outside the US, many scholars nevertheless struggle how to measure judges ideological positions. This is all the more difficult for, say, students of constitutional courts in Europe. Most constitutional courts do not publish individual votes, rendering it simply impossible to estimate judges' ideological positions based on an analysis of observed individual voting patterns (e.g., Bailey and Chang, 2001; Bailey and Maltzman, 2011; Bailey, 2013; Hanretty, 2012; Martin and Quinn, 2002, 2007). Even within EU member states the institution to file a separate opinion is not in the repertoire of actions for constitutional court judges in seven countries (Raffaelli, 2012).

Thus, the strategy of choice for many applications outside the US context is to employ the *party-label heuristic* and simply assign each judge the ideology score of the party (or any other political actor) that nominated or appointed the judge. Based on the distribution of scores across judges, scholars apply an aggregation rule (i.e., mean or median) to generate an ideological position for the panel of judges or the entire court. A leap of faith, however, is the assumption that appointees from different parties behave differently. We cannot test that directly when individual votes are not published.

The contribution of our paper is that we provide a strict test of this widely-held assumption and show that party labels seem to be diagnostic in terms of expected behavior of judges even if we do not observe their individual votes. We closely look at day-to-day operations at the procedural level within the German Federal Constitutional Court (GFCC) that have been seemingly below the radar of many students and observers of the court so far. We argue that there is at least a

minimal effect of ideology if we find that party labels matter already at the procedural level.

In particular, we analyze how the court organizes its work-flow and show that a judge's party-label matters. As many other constitutional courts around the world without full docket control, the GFCC created several three-judges panels, which are called *Chambers*, to be able to make timely decisions for a large number of cases that are deemed to be not important or controversial enough to be deliberated among all judges on the bench.

Given how the judges get nominated and finally elected to the GFCC, applying the *party-label heuristic* we can distinguish two differently labeled judges. Using an original data set of all Chamber compositions and their 3944 published decisions between 1998-2011 we provide two tests that party labels matter. First, we show that so-called "*homogenous*" (unicolored) panels consisting exclusively of judges nominated by the same party can be observed *less often* than they should by chance alone if party labels do not matter.

Second, using a quasi-experimental approach we show that there are systematically more deviations from the court's rule of procedure on how to substitute absentee judges for the composition of the court's Chambers if those rules would produce "unicolored" panels. Given that party labels seem to matter already at the procedural level, we conclude that ideology should also play a role when it comes to decision-making in the Senates.

2 The Minimal Effect of Ideology and the Organizational Structure of the German Federal Constitutional Court

When studying judicial decision-making in situations where judges' individual decisions are not disclosed, scholars cannot use voting patterns to infer from decision outcomes to the ideological positions of judges. A common alternative strategy is to employ the so-called *party-label heuristic* and use the ideological position of the judge's appointing party as a proxy for the ideological positions of the judge. However, such a measurement strategy assumes that judges really do behave according to their party label and do not change their individual preferences over time, as it is known from the US Supreme Court (Martin and Quinn, 2002).

Furthermore, literature has exclusively focused on the final decision output of courts, e.g. whether legislation was rejected or not. Scholars have shown that besides ideological considerations (Hönnige, 2009; Hanretty, 2012), also public opinion (Vanberg, 2005) and legal doctrine (Bailey and Maltzman, 2011) matters. This makes it hard to pinpoint judge's motivation and distinguish between different causal effects that might operate side by side.

In this paper, we show a way to overcome this shortcomings by providing an alternative identification strategy of judge's ideological motives. We argue for a *minimal effect* of ideology,

that is that ideological motives already affect judge's everyday thinking. Our identification strategy is thereby to test the influence of ideology first in a soft-case scenario using data on day-to-day operations, and then transmit our findings to the more politicized final decision level as a hard-case scenario. If we manage to show that even court's daily routine is shaped by partisan struggle amongst judges, yet it is more likely that ideological thinking also condenses in the political environment of Senate decisions.

We show the "minimal effect" of ideological constrained judges using the German Federal Constitutional Court (GFCC) as study subject. The clue of our research strategy is that we investigate the ideological effect already at the procedural level of the court. We chose the German court as it is considered as one of the most influential courts worldwide, and it also functions as a blueprint for many other European courts and courts in transitional countries.

Judges of the GFCC are appointed through a political process that is based on a complex inter-party agreement, where the right of nomination rotates between the two big German parties Social Democrats (SPD), Christian Democrats (CDU/CSU) and the two smaller parties The Greens and the Liberal Democrats (FDP) (Hönnige, 2009). Hence, judges are branded with a "red" (SPD), "black" (CDU/ CSU), "green" (Greens) or "yellow" (FDP) party label[1]. Although the court is supposed to be politically independent, parties are very concerned that no party reaches the majority in a Senate and the "golden rule of parity" is not broken. Thus, the Senate's composition is politically contested but heterogenous.

The GFCC consist of two Senates, each with eight judges. However, the daily work load is mastered by panels of three judges, the so-called *Chambers*. Those panels (each Senate has three of them) account for 98 percent of the total decision volume, which makes them an important yet ignored part of the court. Those panels have a "filter function": they check whether the legal requirements of complains are met, but can also decide on cases without having consulted the Senates if the legal outcome of a case is sufficiently distinct. The three judges of each panel vote unanimously. This makes each panel's judge a veto player.

2.1 First Implication: Composition of Chambers

The decision how to compose three-judge panels allow us already to test a first implication regarding the party-label heuristic. If party labels do not play any role for the decision how to compose them than we should observe at least some homogenous, unicolored panels just by chance. Given that the assignment to various three-judge panels are done within the Senates (eight-judge panel), there are $56(= \frac{8\times7\times6}{1\times2\times3})$ different three-judge panel compositions. Given that among the eight

[1]Throughout this paper we continue to use this "color terminology" for a better readability and illustration of our examples.

judges per Senate there are an equal number of "black" and "red" judges[2] we should observe 4 ($= \frac{4\times3\times2}{3\times2\times1}$) homogenous panels when *sampling without replacement* per Senate. Thus, the probability to observe a homogenous panel of judges just randomly is about 7 % ($= \frac{4}{56}$) per Senate. In fact, across both Senates we observe no (!) "red" chambers and merely 4 "black" chambers among 141 different three-judge panels in our data, i.e., in only less than 3% ($= \frac{4}{141}$) of all the different three-judge panels that have been composed within our observation period. Thus, this evidence already suggests that the court seems to take the party-label of the judges into account when composing different three-judge chambers, and tries to actively avoid homogenous ones. Therefore, we conclude that at least at the level organizational level – below the level of actual decisions – party-label matters. In the reminder of our paper we develop and discuss our second test. We thus turn to the rules of procedure of the court and expect to find systematic deviations from it under certain conditions.

2.2 Second Implication: Deviation fro the Rules of Procedure

The allocation of judges to the panels is defined in the *rules of procedure (RoP)* (*German: Geschäftsverteilung*). These rules are adopted by each of the Senates at the beginning of each year and strictly settle which judge works in which panel. They also codify what happens when judges become indispensable and drop-out of work, e.g. due to sickness, vacations, or other reasons. In those cases, the rules of procedure exactly determine which judge is replaced by whom. Thus, for drop-out cases there exist fixed substitution patterns.

Just as the Senates, the panels and their compositions are ideologically *heterogenous*. This means that there is no Chamber that consists of judges that all have the same party affiliation. Or, in metaphorical terms: no political "color" has a dominating position in a Chamber[3]. Hence, the typical panel compositions look as illustrated in Figure 1 :

Chamber 1			**Chamber 2**			**Chamber 3**		
Red	Red	Black	Red	Black	Red	Black	Black	Red

Table 1: Typical composition of heterogenous Chambers of the FCC. Red stands for SPD-nominated judges, whereas black corresponds to CDU/ CSU nominated ones.

But why does the German court pay so much attention to a heterogenous composition even on the Chamber-, the lowest organizational level? We argue that judges are political animals and that their behavior is driven by ideological considerations. Two mechanisms are conceivable: first, political motives could play a role because decisions of the court have political implications (*direct*

[2]More precisely there is sometimes a "yellow" judge nominated on the CDU/CSU ticket which we code therefore as a "black" judge and, conversely, a "green" judge nominated on the SPD ticket, which we code as a "red" judge.

[3]Media labeled homogenous "red" panels as "Rotlichtkammer" (Redlight-Chamber), whereas homogenous "black" panels are labeled as "Dunkelkammer" (Dark-Chambers).

effect of ideology). There is evidence that judges have political preferences that correspond to the political player that appointed them (Hönnige, 2009), thereby turning judges into "policy-seekers". In order to maximize their own utility, judges seek to avoid that colleagues with another party label dominate a panel and thereby can shape decisions in their favor. For this reason, judges balance the composition of panels already formally. Second, judges could self-restrain because they know their actions are monitored by the public and the media. Although the decision-making of the Chambers is not as present in the media and public's mind as Senate decisions, Chamber decisions happened to be criticized by influential German newspapers. Appearing to be politically biased could reduce the legitimacy of the court's decisions and harm its reputation as an unconstrained court. Whatever mechanism applies, the outcome is the same: Judges have an interest in avoiding panels that are "homogenous", hence behaving as political animals

3 Identification Strategy

As identification strategy of the minimal effect of ideology we apply a quasi- experimental approach, which is defined as a "real- world situation that produces haphazard assignment to a treatment" (Rosenbaum, 2010, 67). Through a natural intervention we seek to create an as- if randomized treatment assignment that achieves independence between treatment assignment and unit level potential outcomes. We use the potential outcomes framework (Rubin, 1974) (known as the Rubin Causal Model, RCM) to formalize our identification strategy.

Written in potential outcome terms, our design looks as follows:

Units (i) = Episodes of Chamber decisions

Outcome (Y) = Deviation from the rules of procedures ($Y_i \in \{0; 1\}$)

Treatment (D) = Critical case ($D_i \in \{0; 1\}$)

Treatment assignment mechanism (Z) = Random drop- out of a judge

In our application to the Chambers of the Bundesverfassungsgericht, we want to explain the binary outcome Y_i *deviation from the rules of procedure* through the binary treatment D_i *critical case* for each episode of Chamber decision i.

By D *critical case* we refer to cases where mechanically following the substitution pattern of the rules of procedure leads to unbalanced Chambers. This is best illustrated using Figure 1: Imagine the rules of procedure require Chamber 1 to replace a sick judge with the last- named judge from

Chamber 2. Thus, a drop- out of the black judge in Chamber 1 and a replacement of him or her through the last- named judge in Chamber 2 (a red judge) would cause Chamber 1 consisting of three red judges (Redlight- Chamber). Hence, Chamber 1 is unicolored and homogenous.

Furthermore, Y describes the phenomenon of judges *deviating from the rules of procedure*. By deviation we mean that judges do not follow the formal replacement order that is defined in the rules of procedure. Illustrating this again using the example from above, imagine that the court skips the last- named (red) judge from Chamber 2, but uses the second- last- named (black) judge as a replacement. Now, Chamber 1 is still heterogenous, as a black judge dropping- out is replaced by another black judge coming in.

The above scenario already reveals the linchpin of our research design. If judges are political animals and want to avoid unbalanced panels, they should deviate from the rules of procedure more often in critical cases than in cases where mechanically following the rules does not distort the Chamber's color composition. As an observable implication, we therefore expect the deviation rate to be higher in critical cases than in non- critical ones.

As credible identification strategy we leverage the fact that judges being absent (e.g. due to illness) and the resulting occurrence of a critical case do not follow systematic patterns. This makes Z *drop- out of a judge* a real- world situation that mimics a randomized experiment. The drop- out thus creates a control group ($D = 0$) and a treatment group ($D = 1$) where treatment D is administered completely random. This allows us to observe a causal effect operating in relative isolation from threats of confounding (Keele, 2015, 319). Because the potential outcomes are Y_{iD}, the actual outcome is a function of treatment assignment and potential outcomes such that $Y_i = D_i Y_{i1} + (1 - D_i)Y_{i0}$ (Keele, 2015).

The as- if randomized nature of our treatment assignment allows us to assess the causal effect by estimating the average treatment effect (ATE) that is a comparison of the difference of the unit- level potential outcomes in mean outcomes for the treated and untreated units i:

$$ATE = \delta = \delta^{naive} = E(Y^1|D = 1) - E(Y^0|D = 0), \text{ or } \alpha_i = Y_{i1} - Y_{i0}.$$

This is because the treatment assignment is independent of potential outcomes $(Y_1, Y_0) \perp D$).

In a second step we want to justify why the nature of our assignment is as good as random. We do this by highlighting several potential confounders in our research design. For a graphical illustration we rely on the directed acyclic graphs (DAGs) framework presented by Pearl (1995, 2009). We first name the confounders and their respective place in the causal framework and then substantiate why our identification strategy is still valid.

Figure 1 shows graphically how potential confounders in our research design are located.

There are two potential (and observable) confounders that are related to the treatment assignment mechanism Z, namely A *Age* and P $Professorship$. Recall that Z was the drop- out

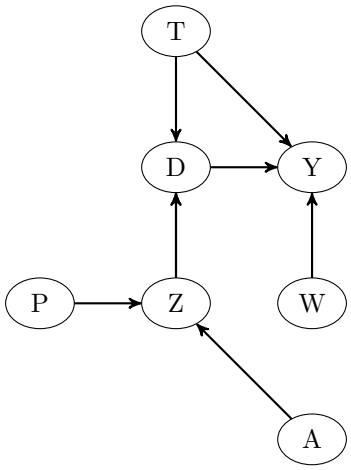

Figure 1: DAG of the effect of *critical cases* D on the *deviation of the rules of procedure* Y. Z is the *random treatment assignment* through sickness, A represents the *age* of judges, whereas P stands for a *professorship* of judges. W stands for the judge's *workload*. T describes a *tactical drop- out*.

of a judge from a chamber- decision, e.g. due to illness. In expectation, older people become more often sick compared to younger ones. This implies different probabilities of dropping out among judges. The same logic accounts for P: beside their mandate as constitutional court judge, some of the court members also hold a professorship at universities (approximately one third of all Bundesverfassungericht judges hold such a professorship). There is the chance that those judges drop- out from court business more often than their colleagues without such a mandate because of professors additional responsibilities to prepare and give lectures or to participate in academic committees.

With respect to the validity of our research design, A and P do no constitute a theoretical threat to our identification strategy. As Figure 1 reveals, A and B do not have any causal relationship to Y. There is no theoretical reason to believe that age or holding a professorship should cause panels to systemically validate the RoP. Hence, as long as older judges or judges with a professorship do not appear systematically more often in panels that deviate significantly more from the RoP, there is no need to block their backdoor path.

W *Workload* describes the individual workload of judges and the consequences for a occurrence of Y. There are eight judges per Senate, but nine judges are needed to occupy the three-judge-panels. As logical consequence, one judge has to be part of two Chambers. Typically the President of each Senate holds this double function. Being part of two Chambers plus being President of a Senate obviously leads to a substantial workload. The GFCC usually accounts for this problem by putting those judges on last place in the replacement order. However, sometimes those judges with a heavy workload appear on place two or even place one of the replacement order. Obviously this harms their ability to function as a replacement for others: the higher their workload and the earlier they are in the replacement order, the higher the chance that they are not on hand

for a replacement, which makes it more likely that there is a rules of procedure deviation (Y). Nevertheless, this is not a problem for our quasi- experimental design: in expectation, workload W has no causal relation on D.

There is also an (unobservable) potential confounder in terms of T $Tactical - drop - out$ of judges. T describes the chance that judges might break out of the random selection procedure by dropping out intentionally and hence causing a deviation from the rules of procedure due to (unknown and unobserved strategic considerations. Thereby the appearance of a critical case is no longer random and we are not able to gain an unbiased estimate of the causal effect of D on Y anymore. Blocking the backdoorpath via T is crucial for our design. However, there is no way for us to observe whether a judge calls in sick strategically or because of a real sickness. Yet there are strong theoretical reasons against such a strategic drop- out behavior: at best, judges would not benefit from it. At worst, judges would help judges with a competing party affiliation. Recall that the panels normally are always balanced politically according to the rules of procedure. Given Chamber 1 from Figure 1, a red judge who calls in sick cannot change the balance in the Chamber, because he or she is either replaced by another red judge (nothing changes) or a black judge (which means that now two black and one red judge are in the panel). Due to unanimity, a heterogenous panel does not benefit any side. If a black judge calls in sick in Chamber 1, this either leads to no changes (if another black substitutes) or the whole panel becomes red (if the black judge is substituted by a red one) and unbalanced. As previously explained, such a red and homogenous panel is something that a black judge seeks to avoid at any cost. As a result, from a strategic perspective neither red nor black judges have any incentive to strategically calling in sick. Therefore, we are convinced that our identification strategy and the randomization mechanism is successful.

4 Data

For our analysis we take into account 13 years of Chamber decisions from the German constitutional court. All in all, we obtained 3944 cases between 01/01/1998 and 12/31/2011 from the website of the German Federal Constitutional Court. Our choosing of this time-period is primarily due to data availability concerns. Prior data in digital format is not available, since Court cases were published online only from 1998 onwards. Note that not all Chamber cases are published, for instance if several cases with same contents are abridged to one main trial. Moreover, judges are free to decide whether a decision shall be published. Thus, a potential selection bias is a viable concern about our data base. However, if judges use their discretionary power over publication selectively, we would image judges rather avoiding to publish cases in which replacement allocation did not follow the rules of procedure, than cases in which the rules of procedure were followed.

This selection bias would work against the hypothesized direction of the relationship. Since this would lead us to obtain a conservative estimate, a potential selection bias does not endanger the viability of our research design.

4.1 Identifying drop outs and replacements

All decisions in our sample are signed by the judges taking part in it. From this we extract the exact composition of the responsible chamber for each decision. We then identify the intended composition defined in the Court's rules of procedure for each decision. For each decision, this yields the set of names of the three defacto judges, and set of the three judges intended by the RoP. The difference between these sets identifies the judges that were replaced and the replacing judges. In 247 out of the 3944 cases, judges were replaced.

In a second step, we identify for each decision the order of replacement of judges as defined by the RoP. As there are three judges in three chambers[4], for each chamber decision the RoP stipulates the order in which the other six judges replace drop outs in this chamber. We amend the order of replacement for logical consistency. First of all, since there are only eight judges in each Senate - one judge is allocated to two chambers - the situation can arise that a judge is either his/her own replacement or is mentioned twice in the order of replacement. In the first instance, we delete the judge in question from the order of replacement. In the second instance, we delete the second mentioning of the judge from the order of replacement. This leaves us with five judges in the order of replacement for each decision. Moreover, we make sure that judges reported to be absent are not counted as potential replacements. We therefore track the absence periods of each judge, and delete this judge during his/her time of absence from the order of replacement of other chamber.

4.2 Identifying replacement episodes

In a third step, we carefully define our unit of analysis, which is the replacement episode, not the individual case. A replacement episode is defined as the period between changes in the replacement pattern. It follows that all cases that have the same replacement pattern are combined into one replacement episode. We opt for the replacement episode because replacement patterns in our data show strong path-dependency. Judges that are absent in multiple consecutive cases are generally replaced by the same replacement judge. This indicates that replacement allocation at the Court is most likely not administered on a case-by-case basis. This makes sense from an organizational perspective. Carrying over past replacement allocations as long as the same judge is absent is an organizational shortcut that helps to keep the administrative workload of

[4]Due to organizational matters, the First Senate had four panels from 01/01/2000 until 10/04/2002.

assigning replacement judges manageable. Cases are therefore not the appropriate level of analysis to study RoP deviations in replacement allocations. Not taking into account the strong path dependency of replacement allocation runs the risk of duplicating units of analysis. There are also statistical reasons for our choice. Our quantity of interest, the probability of a RoP deviation, would not be the result of statistically independent Bernoulli trials if we were to use these duplicate observations. Conducting our empirical analysis on the case level would then require a more complicated analytical strategy that would heavily rely on an assumption-laden model of the dependence between cases. To circumvent these problems, we spent some notable effort to carefully combine cases in the same replacement episode.

For cases with a single absentee judge, what constitutes a replacement episode is relatively straightforward. To facilitate the reader's understanding, Table 2 provides an exemplary account of our coding operations. If subsequent cases within one chamber show the same replacement pattern, i.e. the same absentee judge is replaced by the same replacement judge, all these cases are counted only as one single unit (Cases 1 and 2). If either the absentee judge or the replacement judge changes, or if the absentee judge is back available (Case 3), the replacement episode is terminated and the following cases constitute a new unit of analysis. If the times of absence of two judges overlap (judges C and B in Cases 4, 5, 6), our definition of what constitutes one unit of analysis gets somewhat more complex. If the replacement pattern of two consecutive case is partially the same, only the changes in the replacement pattern are considered. In effect, the prior replacement is treated as given, and only the latter, additional replacement is counted. Cases 4 and 5 serve as an example: while in Case 4 judge D replaces judge C, in Case 5 D and F replaces C and B. Here the subset 'D replaces C' is not counted as part of the replacement episode in Case 5, because the operation 'D replaces C' is assumed to be carried over from the prior replacement episode in Case 4. In effect, judge D is taken as given in Case 5. Only the change in the replacement pattern, that is 'F replaces B', is taken into account. Naturally, cases in which no judge was absent are not counted as replacement episodes (Case 3).

Our combination of one or more cases to replacement episodes decreases our number of observations considerable: Starting with 247 cases in which we noted changes between the formal and defacto composition of the chamber, we are left with only 177 replacement episodes.

4.3 Identifying RoP deviations and critical cases

Finally, we code our dependent and independent variable from the replacement episode-level data. Our dependent variable, *RoP deviation* indicates whether there was a deviation from the rules of procedure in the replacement allocation. This is the case if the order in which replacement judges are allocated does not follow the order of replacement as defined in the rules of procedure. Table 3 provides an example: Replacement episodes 1 and 2 do not constitute RoP deviations,

Case	formal composition	defacto composition	replacement pattern	addition to repl. pattern	replacement episode
1	ABC	ABD	D → C	D → C	1
2	ABC	ABD	D → C	.	.
3	ABC	ABC	.	.	.
4	ABC	ABD	D → C	D → C	2
5	ABC	AFD	DF → CB	F → B	3
6	ABC	AFD	DF → CB	.	.

Table 2: Coding example for identification of replacement episodes

since replacement judge D is first in the order of replacement. In Case 3, judge D is deleted from the order of replacement, since she is already part of the defacto chamber. Judge E is the RoP-conform replacement. But since judge F was allocated as a replacement, this replacement episode constitutes an RoP deviation.

Our independent variable *critical case* indicates whether the (hypothetical) composition of the chamber if the replacement order had been followed, is unbalanced. A chamber is unbalanced if it contains only judges of the same "color", which is indicated by their party affiliation. Judges can have two colors, black and red. Black judges were nominated by rightist parties (CDU, CSU, FDP), red judges by leftist parties (SPD, Greens). Turning to our example in Table 3, in episodes 1 and 2, the RoP-conform composition of the chamber contains two black and one red judge, therefore these are not critical cases. In episode 3, an all-black chamber would have resulted from following an RoP-conform replacement. Therefore episode 3 constitutes a critical case.

repl. episode	formal	defacto	add. to repl. pattern	order of repl.	RoP deviation	composition after RoP-conform repl.	critical case
1	ABC	ABD	D → C	DEFGH	0	ABD	0
2	ABC	ABD	D → C	DEFGH	0	ABD	0
3	ABC	AFD	F → B	~~D~~EFGH	1	AED	1

Table 3: Coding example: Identification of RoP deviation and critical cases

5 Analysis

The causal effect, our quantity of interest, is then the difference in the probability of RoP violation between critical and non-critical cases. First of all, we analyze the relative frequency of RoP deviations. This should already give us a clear indication whether RoP deviations are more likely in critical cases, as the relative frequency is the maximum likelihood point estimate of our quantity of interest. Since the RoP's of two senates specify different replacement rules, the first based on the inverse hierarchy, the second on inverse seniority, we analyze RoP deviations separately for each senate.

As Table 4 indicates, our data shows clear signs that RoP deviations are more likely in critical cases than in non-critical cases, at least in the first Senate. Here the relative frequency with which RoP deviations appeared in non-critical cases is only 26%, while the relative frequency with which replacements deviated from the order specified in the RoP in critical cases was 62%, in 8 our of 13 cases. In Senate 2, the pattern is much less clear. Here RoP deviations appeared in 57% of non-critical cases, and in 62% of critical cases. Although the difference points in the hypothesized directions - with RoP deviations 5 percentage points more likely in critical cases - the magnitude of the observed difference in not substantial.

Senate 1	non-critical case	critical case
no RoP deviation	51	5
	(74%)	(38%)
RoP deviation	18	8
	(26%)	(62%)
Senate 2	non-critical case	critical case
no RoP deviation	32	8
	(43%)	(38%)
RoP deviation	42	13
	(57%)	(62%)

Table 4: Absolute and relative frequency of RoP deviations in critical and non-critical cases.

Given the small sample size, we test whether the RoP violation rate is significantly larger in critical than in non-critical cases. To do so, we run three logit models, one on the combined sample, one on the first and one on the second senate.

Table 5

	RoP deviation		
	Both Senates	Senate 1	Senate 2
Critical case	0.8^{*}	1.5^{*}	0.2
	(0.4)	(0.6)	(0.5)
Constant	-0.3	-1.0^{**}	0.3
	(0.2)	(0.3)	(0.2)
Observations	177	82	95
Note:	*p<0.05; **p<0.01; ***p<[0.***]		

In all three models, the signs of the coefficient point in the predicted direction - RoP deviation is more likely in critical cases. For the combined senates and the first Senate, we find the coefficient to be statistically distinguishable from zero, employing a confidence level of 95%. As expected, the small difference in violation rate between critical and non-critical cases is not statistically significant in Senate 2.

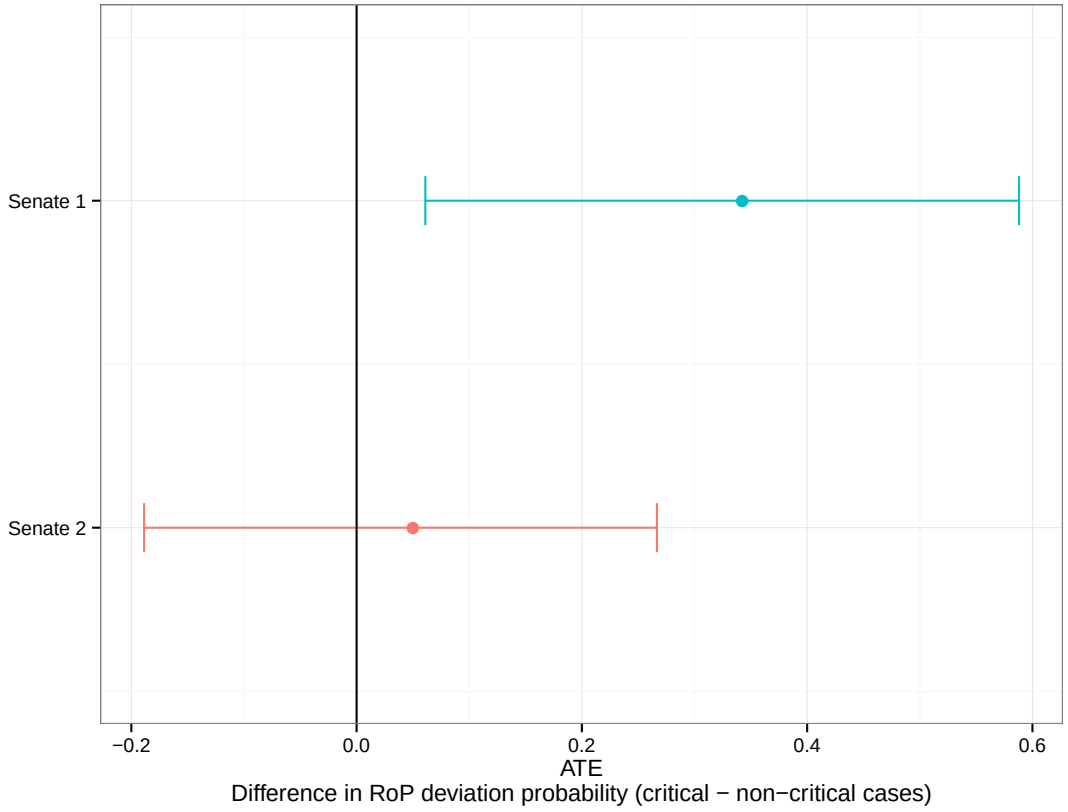

Figure 2: Average treatment effect and 95% confidence intervals by Senates.

From the model estimates, we calculate our quantity of interest, the first difference in RoP deviation probability between critical and non-critical cases, using standard simulation techniques. As shown in Figure 2, in Senate 1, we estimate RoP deviation to be 34 [6,59] percentage points more likely in critical than in non-critical cases. For Senate 2, we fail to establish a statistical significant causal effect (5 [-19,27]).

6 Conclusion

In this paper we argue for a *minimal effect* of ideology on judges. Contrary to current approaches that attempt to demonstrate the impact of party label heuristic on judicial behavior at the highest level of courts, their final decisions, we find evidence that judges' thinking is already governed by party politics at day-to-day operations. Using a novel data set on Chamber decisions of the German Federal Constitutional Court from 1998- 2011, we show that judges systematically deviate from the rules of procedure in order to avoid panels of judges that are nominated by the same party. Our contribution is that we demonstrate the importance of political motives for judges based on a quasi- experimental approach that gets rid of potential confounders already through its research

design.

To find an answer to this question, future work is dedicated to assess the implications of our empirical findings: do "homogenous", unicolored Chambers really make ideologically more extreme decisions? How can we empirical test this implication, given that individual votes of Chambers are not published and standard identification strategies cannot be implemented? And what normative implications do

References

Bailey, Michael A. 2013. "Is Today's Court the Most Conservative in Sixty Years? Challenges and Opportunities in Measuring Judicial Preferences." *The Journal of Politics* 75(3):1–14.

Bailey, Michael A. and Forrest Maltzman. 2011. *The Constrained Court: Law, Politics, and the Decisions Justices Make*. Princeton University Press.

Bailey, Michael A. and Kelly H. Chang. 2001. "Comparing Presidents, Senators, and Justices: Interinstitutional Preference Estimation." *The Journal of Law, Economics & Organization* 17(2):477–506.

Dyevre, Arthur. 2010. "Unifying the field of comparative judicial politics: towards a general theory of judicial behaviour." *European Political Science Review* 2(02):297–327.

Hanretty, Chris. 2012. "Dissent in Iberia: The ideal points of justices on the Spanish and Portuguese Constitutional Tribunals." *European Journal of Political Research* 51(5):671–692.

Hönnige, Christoph. 2009. "The Electoral Connection: How the Pivotal Judge Affects Oppositional Success at European Constitutional Courts." *West European Politics* 32(5):963–984.

Keele, Luke. 2015. "The Statistics of Causal Inference: A View from Political Methodology." *Political Analysis* 23(3):313–335.

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court , 1953-1999." *Political Analysis* 10(2):134–153.

Martin, Andrew D. and Kevin M. Quinn. 2007. "Assessing preference change on the US Supreme Court." *Journal of Law, Economics, and Organization* 23(2):365–385.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82(4):669–710.

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

Raffaelli, Rosa. 2012. Dissenting opinions in the Supreme Courts of Member States. Technical report.

Rosenbaum, Paul R. 2010. *Design Of Observational Studies*. New York: Springer.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5):688–701.

Sunstein, Cass R., David Schkade, Lisa M. Ellman and Andres Sawicki. 2006. *Are Judges Political?: An Empirical Analysis of the Federal Judiciary*. Washington D.C.: Brookings Institution Press.

Vanberg, Georg. 2005. *The Politics of Constitutional Review in Germany*. Cambridge: Cambridge University Press.