Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Multitasking, Multiarmed Bandits, and the Italian Judiciary

Robert L. Bray, Decio Coviello, Andrea Ichino, Nicola Persico

To cite this article:
Robert L. Bray, Decio Coviello, Andrea Ichino, Nicola Persico (2016) Multitasking, Multiarmed Bandits, and the Italian Judiciary. Manufacturing & Service Operations Management 18(4):545-558. http://dx.doi.org/10.1287/msom.2016.0586

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 18, No. 4, Fall 2016, pp. 545–558
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2016.0586

Multitasking, Multiarmed Bandits, and the Italian Judiciary

Robert L. Bray
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, r-bray@kellogg.northwestern.edu

Decio Coviello
HEC Montréal, Montréal, Québec H3T 2A7, Canada, decio.coviello@gmail.com

Andrea Ichino
European University Institute, 50014 San Domenico di Fiesole FI, Italy; and University of Bologna, 40126 Bologna, Italy, andrea.Ichino@eui.eu

Nicola Persico
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, nicola@nicolapersico.com

We model how a judge schedules cases as a multiarmed bandit problem. The model indicates that a first-in-first-out (FIFO) scheduling policy is optimal when the case completion hazard rate function is monotonic. But there are two ways to implement FIFO in this context: at the hearing level or at the case level. Our model indicates that the former policy, prioritizing the oldest hearing, is optimal when the case completion hazard rate function decreases, and the latter policy, prioritizing the oldest case, is optimal when the case completion hazard rate function increases. This result convinced six judges of the Roman Labor Court of Appeals—a court that exhibits increasing hazard rates—to switch from hearing-level FIFO to case-level FIFO. Tracking these judges for eight years, we estimate that our intervention decreased the average case duration by 12% and the probability of a decision being appealed to the Italian supreme court by 3.8%, relative to a 44-judge control sample.

Keywords: multitasking; multiarmed bandits; field experiment; production scheduling; Italian judiciary
History: Received: July 17, 2015; accepted March 24, 2016. Published online in *Articles in Advance* August 29, 2016.

1. Introduction

The Italian judiciary is slow. The World Bank ranks Italy 147th out of 189 countries in ease of enforcing contracts: it takes an estimated 3.25 years to enforce a contract in Italy, slightly less than Djibouti (3.35), and slightly more than Myanmar (3.18) (World Bank Group 2014). Among developed countries, Italy is a judicial outlier—twice as slow as any other member of the Organization for Economic Co-operation and Development (OECD), a trade confederation of 34 industrialized countries (OECD 2013). And the problem is getting worse; the stock of pending civil cases increased by 10% from 2008 to 2010 (Esposito et al. 2014, p. 6), and the average Italian civil case duration increased by 19% from 2010 to 2012 (CEPEJ 2014, p. 200). The International Monetary Fund (IMF) argues that the inefficiency of Italian courts leads to "reduced investments, slow growth, and a difficult business environment" (Esposito et al. 2014, p. 1).

We study the Italian Labor Court of Appeals. Italy's appellate courts are especially sluggish: The average Italian case is 2.4 times as long as the average OECD case at the trial level, and 4.7 times as long at the appellate level (OECD 2013). And Italy's labor courts are vital, as the IMF explains, "[Italy's] inefficient labor courts can have detrimental effects on the composition of employment and labor market participation. [They] also affect job reallocation, which in turn impacts productivity and capital intensity" (Esposito et al. 2014, p. 5).

Several top-down initiatives have failed to reform the Italian judiciary. The European Commission for the Efficiency of Justice (CEPEJ) recommended "good practices and innovative suggestions" (CEPEJ 2006), but the Italian judiciary ignored them. The Parliament passed a law that reduced judge summer vacations from 45 to 30 days, but the judiciary refused to apply it (Chirico 2013, ANSA 2015). And the Italian judiciary's self-governing body appointed a special commission to develop productivity benchmarks, but the judges on the panel failed to find consensus (Unita' per la Costituzione 2015).

Italian judges could resist these top-down reforms because they are politically independent and operate with impunity. Accordingly, bottom-up reforms,

shepherded by judges, have more potential for efficacy. To galvanize judicial action, a policy should (i) be clearly beneficial, (ii) preserve judicial autonomy, (iii) not increase workloads, and (iv) be easy to understand and implement. We identify such a policy improvement in the Roman Labor Court of Appeals.

This court's cases generally require two or three hearings. Traditionally, the judges scheduled one at a time, arranging a case's $(n+1)$th hearing during its nth hearing. Consequently, every hearing joined the end of a judge's work queue (when his or her calendar was next free); thus, cases comprising N hearings would cycle through the work queue (the calendar) N times. We call this policy, which prioritizes the oldest hearing, hearing-level FIFO. We propose a new scheduling policy in which cases traverse the docket only once: case-level FIFO. Under this policy, a judge selects the oldest case and works on it to fruition before opening a new case. More accurately, we propose a relaxation of case-level FIFO, as the policy in its strictest sense violates scheduling constraints. The judges implement relaxed case-level FIFO by estimating the number of hearings a new case will require and scheduling that many up front (reserving preparation time between hearings). This scheduling policy mimics case-level FIFO by minimizing queue re-entry.

Switching from hearing-level FIFO to relaxed case-level FIFO decreases the number of cycles through the queue, but increases the length of the queue (as multiple hearings per case queue up). With a multiarmed bandit production scheduling model, we show that the former effect dominates the latter, for an overall flow time drop, when the likelihood of finishing a case increases with the hearing number—i.e., when the case completion hazard rate function increases. Intuitively, when the case completion hazard rate function slopes upward, judges should prioritize cases they have already seen because those cases are more likely to reach completion; this is what case-level FIFO does.

We test this theoretical result empirically with a difference-in-difference research design. First, we show that the Roman Labor Court of Appeals exhibits increasing case completion hazard rates. Second, we assign six judges to a treatment group and 44 judges to a control group. Third, we compel the treated judges to adopt the relaxed case-level FIFO policy. And finally, we measure how the treated judges' operational performance changes from the five years preceding our intervention to the three years following, relative to the control sample.

Before our intervention, the treated and control judges completed cases at the same rate; after our intervention, the treated judges outpaced the control judges by 0.07 cases a day (11%). By horizon end, the treated judges decreased their inventories by 87 cases relative to the control judges, and their case flow times by 111 days (12%). Also, after adopting case-level FIFO, the treated judges' rulings were appealed 3.8% less often relative to those of the control judges, which suggests an improvement in ruling quality.

2. Multitasking Literature Review

Our intervention decreases the degree of judicial multitasking—reducing the number of open cases that judges need to juggle. The effect of reducing multitasking is complex, and the literature has identified multiple pros and cons.

2.1. Prioritization by Service Time

This work's closest antecedents are the judicial multitasking studies of Coviello et al. (2014, 2015). Coviello et al. explain that juggling many cases distracts judges from prioritizing cases that require little remaining service, which leads cases to linger longer in the docket. For example, suppose a judge has two cases, each requiring two hearings. When the judge finishes the first case before starting the second, the average case finishes after $(2+4)/2=3$ hearings, but when the judge switches between the cases, the average case finishes after $(3+4)/2=3.5$ hearings. Multitasking increases the average wait by diverting the judge away from the most pressing case—the one about to finish.

Coviello et al. test this theory in the Labor Court of Milan (a court different from ours). They measure the causal effect of multitasking on case durations with instrumental variables regressions, instrumenting for case juggling with case difficulty. They estimate that increasing judicial multitasking by 1% increases case flow times by 2%.

We refine the analysis of Coviello et al. First, we derive the intuition that judges should prioritize nearly finished cases from a different model; ours frames courthouse scheduling as a stochastic multiarmed bandit problem, whereas theirs frames it as a deterministic fluid approximation. Second, whereas Coviello et al. study multitasking passively, exploiting natural variations in preexisting data, we do so actively, designing a field experiment to isolate its causal effect—whereas they provide an econometric model that suggests judges *could* reduce case flow times, we *actually* reduce case flow times. This distinction is meaningful, because Coviello et al. never explained whether the necessary scheduling changes were feasible in the wild. This oversight enabled detractors in the Italian judicial community to dismiss the findings of Coviello et al., claiming that it would be impossible for them to juggle fewer cases. Our field experiment disproves these critics, explicitly showing that they can decrease flow times by decreasing the degree of multitasking.

2.2. Setup Times

Most frame the cost of multitasking in terms of setup times. For example, Wang et al. (2015) show that physicians at Northwestern Memorial Hospital zigzag between cases, incurring a setup with each deviation. Wang et al. estimate that if they could eliminate setups, the hospital could serve 20% more patients. Batt and Terwiesch (2012, p. 7) likewise find "switching costs increase with increased levels of multitasking" in a hospital: they estimate that increasing the number of patients from the lower quartile to the upper quartile increases delays by 26%.

We consider a different sort of setup cost: forgetting. It would usually take a judge around nine months to return to a case, during which time he or she would see upward of 500 others. Invariably, the judge would forget the original case, and would thus have to spend much of the follow-up hearing reviewing his or her notes. This is a setup cost. But this differs from a traditional production setup cost, as the amount forgotten increases with the time the case lies fallow. Our scheduling policy mitigates judicial forgetting by decreasing the time between hearings from nine months to six weeks.

2.3. Avoiding Idleness

Aral et al. (2012, p. 851) and Kc (2014, p. 168) likewise document multitasking setup costs. But they also report a benefit to multitasking: switching between tasks enables servers to "utilize lulls in one project to accomplish tasks related to other projects"; indeed, "Switching to a new task rather than idly waiting on a pending task can thus increase worker utilization and improve overall productivity." Accordingly, Aral et al. and Kc recommend a modest level of multitasking to minimize setup costs while avoiding idleness. Our case scheduling policy indeed specifies a moderate degree of multitasking. Lawyers need at least six weeks between successive hearings of a case, so our policy maintains six weeks' worth of open cases.

2.4. Worker Motivation

Tan and Netessine (2014) and Staats and Gino (2012) document a second reason to multitask: Switching tasks periodically increases worker morale. Tan and Netessine find that waiters are more focused when assigned more tables, and Staats and Gino find that bankers are more productive when assigned work that varies across days.[1] Changing the judges' scheduling policies from hearing-level to case-level FIFO is unlikely to influence the judges' motivation because they never see a case twice in one month under either scheduling policy.

[1] Tan and Netessine's and Staats and Gino's findings could also stem from unrelated workload effects (Kc and Terwiesch 2009, Powell et al. 2012, Freeman et al. 2016).

3. Theoretical Motivation

We now model a judge's case scheduling decision as a multiarmed bandit problem (Gittins et al. 2011). Operations researchers have used multiarmed bandit models in assortment planning (Caro and Gallien 2007), production scheduling (Pinedo 2012), labor hiring (Arlotto et al. 2014), queuing (Niño-Mora 2012), and revenue management (Mersereau et al. 2009). Our model suggests that hearing-level FIFO—the court's current scheduling policy—is optimal when (i) previously worked-on cases are less likely to finish than new cases and (ii) judges do not forget case facts. Conversely, the model suggests case-level FIFO—our proposed scheduling policy—is optimal when (i) previously worked-on cases are more likely to finish than new cases and (ii) judges do forget case facts. We show that the case-level FIFO optimality conditions hold in this court, motivating our field experiment.

3.1. Model Overview

A judge has a docket of N cases. Each case comprises a random number of hearings. The judge holds one hearing per period. A case's *hearing number* is its number of completed hearings. A case is *open* if its hearing number is positive (i.e., if the judge has held its first hearing). A case's *hearing age* is zero if it has yet to open and otherwise is the number of periods that have elapsed since the case's last hearing. The likelihood of the judge's finishing a case with hearing number n and hearing age a in the next period is $h(n, a)$. We call h the case completion hazard rate function. The judge incurs waiting cost ω for each unfinished case in each period. The judge seeks to minimize the expected discounted waiting cost, discounting at rate $\beta \in [0, 1)$.

3.2. Hearing-Level FIFO Optimality Conditions

The judges in our Italian court currently follow a hearing-level FIFO policy, prioritizing the case with the largest hearing age. This policy is optimal when $h(n, a)$ is decreasing in its first argument and constant in its second. Hazard rates are constant in the hearing age when the judge has a perfect memory. And hazard rates decrease in the hearing number when, for instance, every case has an unobserved type, either "easy" or "hard"; in this setting, every hearing that does not finish the case increases its conditional likelihood of being a hard type.

When hearing age doesn't influence hazard rates, we can model the judge's decision as a classic multiarmed bandit problem. To do so, we reframe the judge's decision as an equivalent reward maximization problem. The judge receives payoff $\omega/(1-\beta)$ every time a case finishes, and seeks to maximize the expected discounted payoff. Following the classic multiarmed bandit solution, the judge works on the

case with the largest Gittins Index; a case with hearing number n has Gittins Index

$$g(n) = \frac{\omega}{1-\beta} \max_{\tau>0} \left(\sum_{t=1}^{\tau} \beta^t h(n+t,0) \cdot \prod_{s=1}^{t-1} \frac{1-h(n+s,0)}{\sum_{t=1}^{\tau} \beta^t \prod_{s=1}^{t-1}[1-h(n+s,0)]} \right).$$

It's straightforward to show that g decreases in the hearing number when h does (Pinedo 2012, p. 278). So the case with the fewest number of completed hearings has the largest Gittins Index, and thus is most deserving of the judge's attention. In this scenario, the judge cycles through the cases, arranging cases by hearing age.

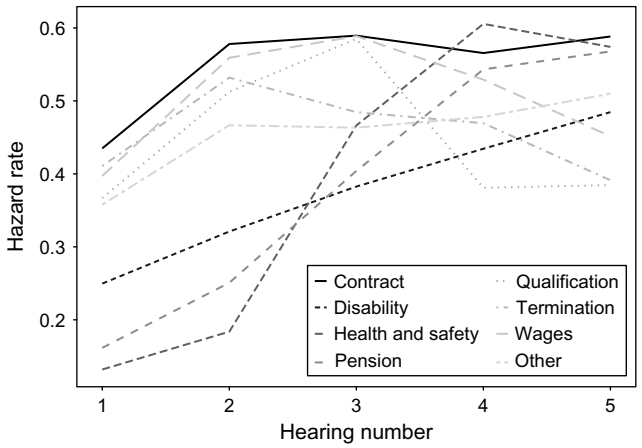
3.3. Case-Level FIFO Optimality Conditions

Our experiment changes the judges' scheduling policy from hearing-level FIFO to case-level FIFO. Case-level FIFO is optimal when $h(n,a)$ is increasing in its first argument and nonincreasing in its second. Hazard rates decrease in the hearing age when the judge has an imperfect memory. And hazard rates increase in the hearing number when, for example, there is a set amount of work that needs to be done; in this setting, every hearing that does not finish the case decreases the expected amount of remaining work.

First, we consider the case in which $h(n,a)$ increases in its first argument and remains constant in its second. In this setting, the Gittins Index solution holds in Section 3.2. But now g *increases* in the hearing number. So the case with the most number of completed hearings has the largest Gittins Index and thus is most deserving of the judge's attention. In this scenario, the judge sees a case through to fruition before starting another.

Second, we show that this case-level FIFO policy remains optimal when the hazard function does not increase in the hearing age. Consider two hazard rate functions h_1 and h_2, where for all n and a: $h_1(n+1,0) \geq h_1(n,0)$, $h_2(n,0) = h_1(n,0)$, $h_1(n,a) = h_1(n,0)$, and $h_2(n,a+1) \leq h_2(n,a)$. First, case-level FIFO is optimal under h_1 because it increases in the hearing number and remains constant in the hearing age. Second, the expected discounted waiting cost is never smaller under h_2 than h_1 because the likelihood of finishing a case is never larger under h_2. Third, the expected discounted waiting cost under case-level FIFO is the same under h_1 and h_2 since the probability of finishing a case is the same under both hazard rate functions when the judge sees cases through to completion. These three claims imply case-level FIFO is optimal under h_2.

Figure 1 Hazard Rates



Note. This plot depicts the case completion hazard rate function by case type.

3.4. Establishing Case-Level FIFO Optimality Conditions

To motivate our field experiment, we demonstrate that the court we study exhibits our two case-level FIFO optimality conditions.

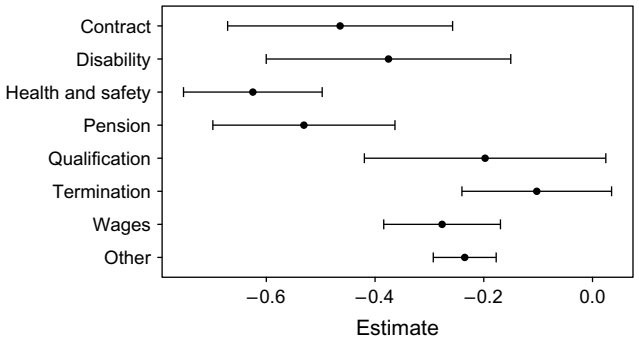
First, we find that the case completion hazard rate function increases in the hearing number. The mean hazard rate—the fraction of hearings that complete a case—increases from 0.29 in the first hearing to 0.37 in the second, to 0.45 in the third, and to 0.50 in the fourth. Each of these increases is significant at the $p = 0.01$ level. And this pattern holds across case types: see Figure 1.[2]

Second, we find that the case completion hazard rate function decreases in the hearing age. Estimating the causal effect of hearing age on case completion hazard rates requires care because the causality can run the other way; high hazard rates can decrease hearing ages when judges prioritize nearly finished cases. To account for this endogeneity, we use two-stage least squares (2SLS), instrumenting for the hearing age with the number of cases on the docket at the time of the previous hearing. (Longer work queues lead to longer interhearing times.)

In our IV regressions, the hearings comprise the observations. (We disregard the first hearing of each cases, which have zero hearing ages.) The dependent variable is $100c_h$, where c_h is a dummy variable indicating whether hearing h completed a case (we scale the dependent variable by 100 to express the hazard rate in percent). The independent variables are the hearing age—the number of days since the case was previously heard—and dummy variables for the month, judge, defendant type, and plaintiff type. The

[2] The hazard rate function decreases after the third hearing in three of the eight case types. But this is negligible, as few cases make it to the fourth hearing.

Figure 2 Effect of Hearing Age on Case Completion Hazard Rates



Notes. This plot exhibits 2SLS regression coefficients. We run the regressions separately for each case type. Our regressions treat each hearing as an observation. Our dependent variable is 100 times a dummy variable that indicates whether the hearing completed a case (we multiply by 100 to express the case completion hazard rates in percent). Our independent variables are the hearing age—the number of days since the case was filed or heard—and dummy variables for the month, judge, defendant type, and plaintiff type. Our instruments are the number of cases on the judge's docket when the given case was last heard or filed, the number of cases on the judge's docket squared, and the month, judge, defendant type, and plaintiff type dummy variables. The points denote the hearing age coefficient estimates: e.g., increasing the hearing age by one day decreases a "Health and safety" case completion hazard rate by 0.76%. And the error bars are the estimates' 95% confidence intervals, derived from robust month-judge block bootstrap standard errors: all but the "Qualification" and "Termination" estimates are significantly negative.

instrumental variables are the number of cases on the judge's docket when the given case was last heard or filed, the number of cases on the judge's docket squared, and the month, judge, defendant type, and plaintiff type dummy variables.

Figure 2 plots the hearing age regression coefficients by case type. The estimates are significantly negative in six out eight case types; judges are less likely to finish cases they haven't seen in a while. This makes sense, as judges must forget case facts over time—it is impossible to perfectly recall 450 cases. For each case type, an F test rejects the null hypothesis of weak instrumental variables at $p = 0.01$ (Stock and Yogo 2005); and for each case type besides "Termination," a Durbin-Wu-Hausman test rejects the null hypothesis that the hearing ages are exogenous at $p = 0.01$ (Davidson and Mackinnon 2004, p. 237).

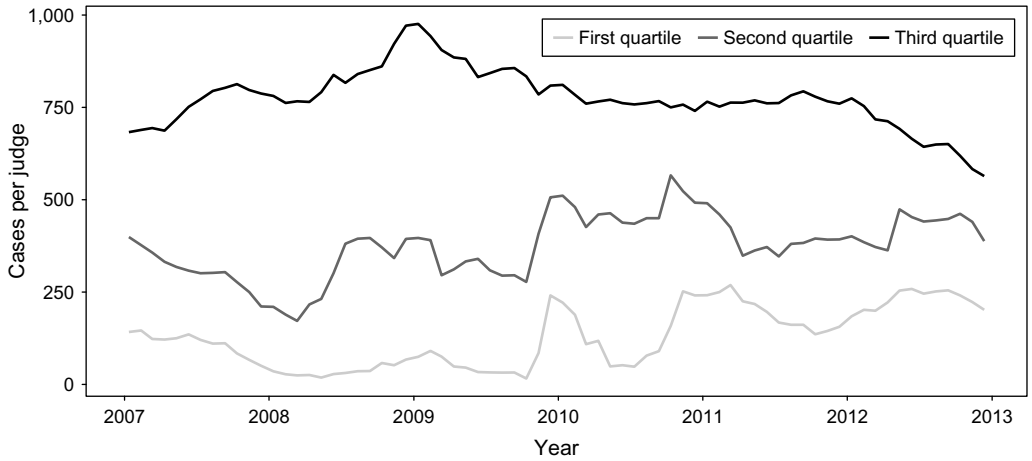
4. Field Experiment

Our theory suggests the Roman Labor Court of Appeals has been implementing FIFO along the wrong dimension. According to our model, the court should follow case-level FIFO, the optimal policy when judges forget case facts and the likelihood of finishing a case increases with the amount of prior work. But it has been following hearing-level FIFO, the optimal policy when judges *never* forget case facts and the likelihood of finishing a case *decreases* with the amount of prior work. We will now test this theory with a field experiment that measures the effect of switching from hearing-level FIFO to case-level FIFO. We use a difference-in-difference research design: six treated judges switch from hearing-level to case-level FIFO on January 1, 2011, and 44 control judges follow hearing-level FIFO throughout.

4.1. Setting

We measure the effect of switching from a hearing-level FIFO to case-level FIFO in the Appellate Labor Court in Rome. This court has jurisdiction over labor appeals in Italy's Lazio Region. It has a long backlog of open cases (see Figure 3). The average case requires 2.3 hearings. The average hearing lasts 20 minutes; the judges can hold upward of 40 hearings a day, rapidly switching between cases.

Figure 3 Inventory of Open Cases



Note. This plot depicts the quartiles of the number of open cases per judge, calculated daily.

The court comprises five *collegios*; one *collegio* comprises our treated judges and the other four our control judges. Each *collegio* contains several three-judge panels. Each case is assigned to a panel for adjudication, and to a *rapporteur*, a judge on the panel, for supervision. The *rapporteur* analyzes the testimony, oversees the ruling, writes the opinion, and schedules the hearings. Since panels are stable, *rapporteur* fixed effects capture panel-level idiosyncrasies; thus, we treat a case's *rapporteur* as its sole judge. The judges hold law degrees, pass selective examinations, hold lifetime appointments, and rarely move between panels.

To minimize corruption—judge shopping and influence peddling—cases are assigned randomly. This random assignment suggests that differences in docket compositions do not drive our results. Indeed, the treated and control samples appear nearly identical in Table's 1 summary statistics.[3] The table provides several insights. First, over a third of cases involve three or more parties; adding a third party increases a case's expected duration by 0.40 hearings, and adding a fourth does so by an additional 0.53 (both of these increases are significant). Second, the conflicts adjudicated are varied; the most common case type category is "Other." Finally, the most common party configuration is a person suing the government (37% of the sample), then a person suing a company (16%), then a company suing a person (11%), and then the government suing a person (8%).

Figure 4 demonstrates that these variables vary over time, so the court is in a constant state of flux. The workloads increase significantly—the average judge receives 16.7 new cases a month before our intervention and 27.1 after—and the judges work faster—the average judge hears 43.1 hearings a month before our intervention and 67.9 after. The case compositions also change; the average number of parties involved drops, as do the defendant and plaintiff type dispersions. These temporal trends motivate our difference-in-difference research design: we cannot naïvely compare the treated judges' pre- and postintervention subsamples because the court is dynamic. But since the treated and control subsamples move in tandem (we will show this more formally in Section 6.2) we can wash out the trends by benchmarking one to another.

Table 1 Summary Statistics in the Preintervention Sample

	Treated		Control		Difference	
	Est.	S.E.	Est.	S.E.	Est.	S.E.
Arrival rate						
Lower quartile	41.00	7.25	36.00	3.31	5.00	7.53
Median	129.50	19.93	90.00	7.03	39.50	19.97
Upper quartile	364.00	42.51	237.00	17.31	127.00	44.11
Service rate						
Lower quartile	138.25	19.81	102.00	8.07	36.25	19.51
Median	433.50	64.89	309.00	19.49	124.50	67.29
Upper quartile	829.00	68.54	635.00	36.32	194.00	84.23
Party count						
Two	58.13	0.56	57.75	0.37	0.38	0.54
Three	19.62	0.36	19.23	0.20	0.38	0.41
Four	9.98	0.20	9.64	0.16	0.34	0.17
Five	5.34	0.16	6.27	0.12	−0.94	0.17
Outcome						
Judgement	85.70	0.37	85.82	0.23	−0.12	0.35
Withdrawl	13.66	0.38	13.47	0.22	0.19	0.35
Settlement	0.64	0.04	0.71	0.02	−0.07	0.04
Case type						
Contract	11.97	0.40	12.03	0.19	−0.06	0.36
Disability	7.73	0.19	7.84	0.12	−0.10	0.22
Health and safety	19.93	0.48	19.70	0.29	0.23	0.46
Pension	7.15	0.20	7.40	0.12	−0.25	0.22
Qualification	1.43	0.07	1.49	0.04	−0.06	0.07
Termination	3.47	0.09	3.57	0.06	−0.10	0.11
Wages	17.10	0.39	16.21	0.21	0.89	0.46
Other	31.22	0.53	31.77	0.25	−0.55	0.55
Defendant type						
Individual	30.48	0.43	29.68	0.25	0.80	0.39
Private company	18.95	0.43	18.31	0.25	0.63	0.45
Public body	44.90	0.75	45.68	0.44	−0.78	0.71
Union	0.33	0.04	0.26	0.01	0.08	0.04
Other	5.34	0.16	6.07	0.09	−0.74	0.15
Plaintiff type						
Individual	65.82	0.41	67.38	0.23	−1.56	0.39
Private company	16.85	0.32	15.48	0.19	1.37	0.27
Public body	10.59	0.22	10.90	0.12	−0.31	0.23
Union	0.27	0.03	0.30	0.02	−0.03	0.04
Other	6.48	0.19	5.94	0.09	0.54	0.21
Hearing count						
One	39.38	2.02	33.94	1.27	5.43	2.53
Two	27.26	1.27	25.24	0.57	2.01	1.43
Three	16.55	0.63	16.48	0.38	0.07	0.83
Four	9.65	0.46	10.74	0.31	−1.09	0.57
Five	4.00	0.23	5.84	0.16	−1.83	0.25

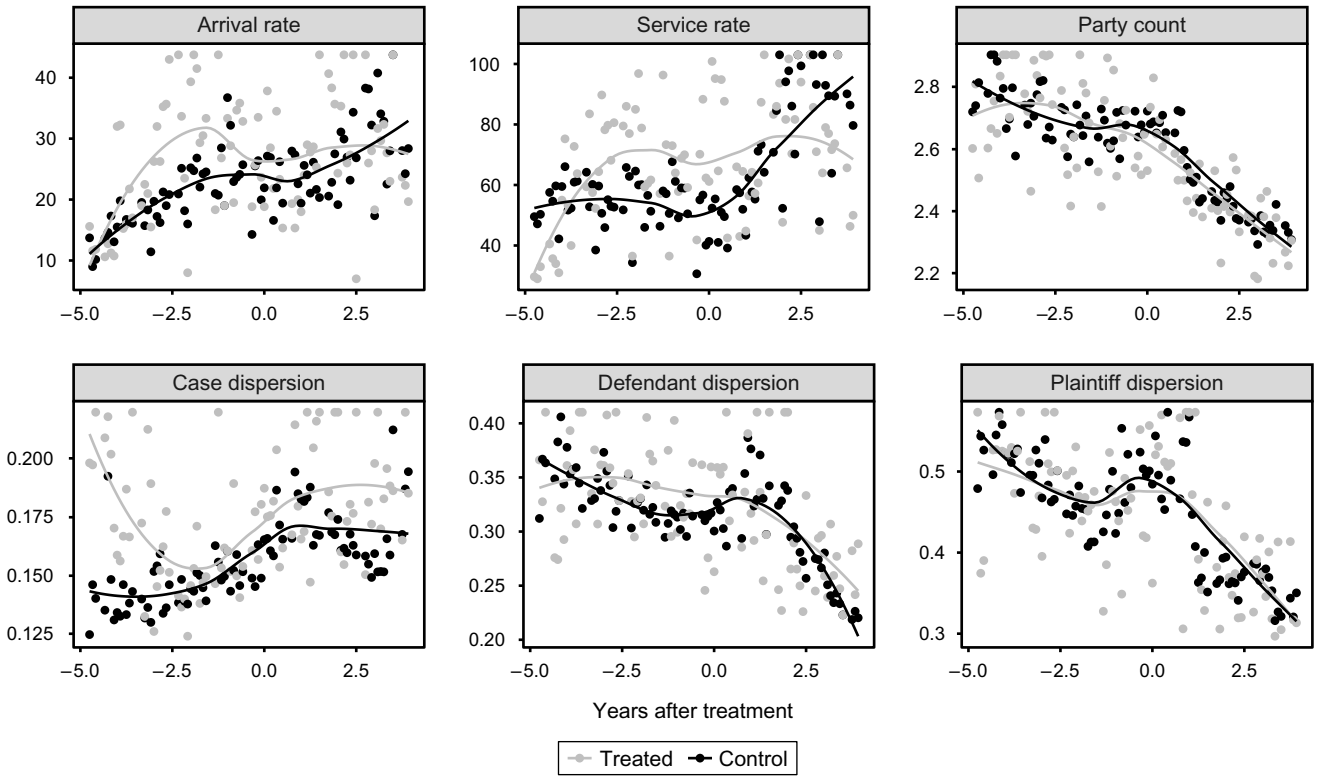
Notes. This table gives the quartiles of (i) case arrival rates measured by cases filed per month per judge and (ii) service rates measured by hearings held per month per judge. It also gives the distribution of (i) the number of parties, (ii) the case outcomes, (iii) the case types, (iv) the defendant types, (v) the plaintiff types, (vi) and the number of hearings required to complete a case. We calculate estimates and standard errors for the treated subsample, the control subsample, and the difference. We only use data from our preintervention subsample.

4.2. Implementation

A president of one of the *collegios* facilitated this experiment. She heard the results of Coviello et al. (2015) at a judicial workshop held in the first instance Court of Rome on October 29, 2009, and emailed us a few months later about the possibility of switching her *collegio*'s two panels to case-level FIFO. We received authorization to conduct the experiment in April 2010 and designed our study the following summer. The *collegio* president convinced her five

[3] The control judges are assigned to fewer cases because some of them have other administrative duties.

Figure 4 Temporal Trends



Notes. This plot exhibits six time series from our treated and control subsamples. The arrival rate series report the average number of new cases filed per judge; the service rate series report the average number of hearings held per judge; the party count series report the average party count for newly filed cases; and the case, defendant, and plaintiff dispersion series report the Herfindahl indices of the newly filed cases' case type, defendant type, and plaintiff type variables. We sample the series with monthly frequency and calculate the corresponding fitted curves with locally weighted polynomial regression.

constituent judges to adopt a case-level FIFO policy on January 1, 2011. These six judges comprise our treated sample and the 44 judges in the other four *collegios* comprise our control sample.

Our intervention was minimal. We simply explained to the president of the treated *collegio* why and how to implement case-level FIFO. And she, in turn, relayed this information to the five other treated judges (whom we did not meet). Since the treated judges negotiated all scheduling details themselves, our field experiment tests whether they have the wherewithal and inclination to improve their schedules.

A strict case-level FIFO policy is infeasible, however, because the judges must (i) schedule hearings at least two months in advance to accommodate the lawyers' schedules and (ii) space hearings at least six weeks apart to leave the lawyers enough preparation time. Accordingly, we recommended a relaxed case-level FIFO policy. When a new case arrives, the judge estimates the number of hearings it will require and preschedules that many up front, spacing the hearings at least six weeks apart. Scheduling multiple hearings at once clusters them in time so that a case finishes soon after it begins, in accordance with case-level FIFO. To avoid idleness, the judges erred on the side of scheduling too few hearings rather than too many. (They usually scheduled between two and four, depending on case complexity.) When they ran out of time slots, they added new hearings to the end of the queue, as they had done previously.

We measured our intervention's effect by tapping into the Roman Appellate Labor Court's database. The court's clerks input data for every case filed between July 7, 2005, and December 31, 2014: the hearing dates, judge, case type, defendant type, plaintiff type, number of parties, whether the judgment was appealed to the supreme court, and whether the case was settled, abandoned, ruled upon, or still open. The observations from July 7, 2005, to December 31, 2010, comprise our preintervention sample, and the observations from January 1, 2011, to December 31, 2014, our post-intervention sample.[4] Table 2 provides summary statistics.

[4] We only have partial data for hearings after May 31, 2014. We know when they took place but not whether they completed a case. Accordingly, we use these observations in the flow time regressions in Section 5.3 but not in the hazard rate regressions Section 5.1.

Table 2 Summary Statistics

	Treated		Control	
	Before	After	Before	After
Judges	6	6	44	36
Court dates	464	199	1,112	845
Cases	8,677	6,674	35,443	43,168
Hearings	17,243	11,822	72,914	83,385

Note. This table records the distinct number of judges, court dates, cases, and hearings of the treated and control subsamples before and after our intervention.

5. Results

In this section, we report the effect of switching from hearing-level FIFO to case-level FIFO. Both the control and treated judges appear less efficient in the post-intervention subsample because the entire Italian judiciary got slower across our sample horizon (Coviello et al. 2012, Esposito et al. 2014, CEPEJ 2014). But the treated judges are faster than they would have been had they tracked the controls. Specifically, we estimate that our new scheduling policy (i) increased the hazard rate of case completion, (ii) decreased the inventory of open cases, (iii) decreased the case flow time, and (iv) decreased the rate at which the judges' rulings were appealed to the supreme court.

5.1. Hazard Rate Increase

Since case arrival rates are fixed, the only way switching from hearing-level FIFO to case-level FIFO can decrease flow times is by reducing the inventory of open cases. To transition from high- to low-inventory regimes, case outflows must temporarily exceed case inflows. Thus, whereas a scheduling policy change will not affect *long-run* case completion rates, which track the exogenous arrival rates, it should increase *short-run* completion rates as the firm burns through excess stock.

There are only two ways to increase the case completion rate: increase the service rate—the number of hearings per day—or increase the hazard rate of case completion—the likelihood of a given hearing concluding a case (i.e., the ratio of cases completed to hearings held). Since our intervention cannot influence the service rate, which is independent of case sequencing, it must reduce inventories via the hazard rate. Thus, the hazard rate of case completion mediates our intervention's effect. The only way switching to case-level FIFO can decrease flow times is by moving high-hazard hearings—those likely to finish a case—to the front of the queue.

We establish that our intervention increased the treated judges' hazard rates with difference-in-difference regressions. We consider four statistical models, regressing with random and fixed effects, and with and without controls. For the control-free random effects specification, we regress $100c_h$ (see Section 3.4) on four variables: (i) the constant 1 (for the *Intercept*); (ii) *Post*, a post-intervention dummy variable; (iii) *Treated*, a treated judge dummy variable, and (iv) the product of *Post* and *Treated*. For the regressions with controls, we include case type, plaintiff type, defendant type, and party count dummy variables. And for the regressions with fixed effects, we include month and judge dummy variables (which make all but the interaction term redundant).

We calculate standard errors with the bootstrap (Horowitz 2001), specifically the block bootstrap, resampling the data by month-judge to make our standard errors robust to cross-correlations within these clusters (Berkowitz and Kilian 2000, Hardle et al. 2003). Our bootstrapped standard errors are about twice as large as classical alternatives.

Table 3 presents the regression coefficients. The control sample's mean hazard rate is 38.7% before intervention and $38.7 - 7.9 = 31.8\%$ after; the treated sample's hazard rate is $38.7 - 0.36 = 38.4\%$ before intervention and $38.7 - 0.36 - 7.9 + 9.0 = 39.5\%$ after. On average, the treated hazard rate is 1% smaller than the control hazard rate before intervention and 28% larger after.

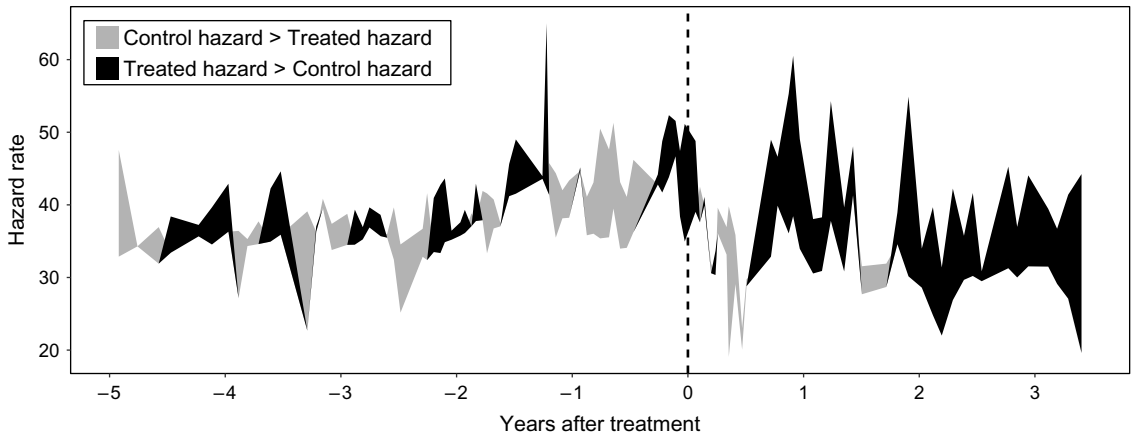
The *Post* · *Treated* interaction term, estimated with fixed effects and without controls, suggests that switching to case-level FIFO increased the treated judges' average hazard rate by 4.85%. In other words, the new scheduling policy enabled the judges to finish an extra case every $100/4.85 = 20.6$ hearings; the

Table 3 Hazard Rate Difference-in-Difference Regression Coefficients

	Without fixed effect		With fixed effect	
	Without controls	With controls	Without controls	With controls
Intercept	38.71	40.28	—	—
	(0.64)	(6.78)	—	—
Post	−7.93	−6.52	—	—
	(0.72)	(0.76)	—	—
Treated	−0.36	−0.23	—	—
	(1.04)	(1.12)	—	—
Post · *Treated*	9.04	8.37	4.85	5.29
	(1.39)	(1.48)	(1.75)	(1.59)

Notes. This table presents ordinary least squares (OLS) regression coefficients. Each observation corresponds to a hearing. The dependent variable is 100 times c_h, a dummy variable that indicates that the hearing concludes a case (we scale the dependent variable by 100 to express hazard rates in percent). The independent variables are (i) the constant 1, (ii) *Post*, a dummy variable indicating the post-intervention subsample, (iii) *Treated*, a dummy variable indicating the treated subsample, and (iv) the interaction of *Post* and *Treated*. The specifications with fixed effects include month and judges dummy variables; the specifications with controls include case type, plaintiff type, defendant type, and party count dummies. We report robust month-judge block bootstrap standard errors in parentheses.

Figure 5 Hazard Rate Time Series



Notes. This plot depicts time series of the treated and control judge hazard rates. We divide the data into 100 time buckets, each of which comprises 1% of the treated judge hearings. We then calculate each bucket's hazard rates with the fraction of cases completed to hearings held. The band's color indicates which sample of judges has a larger hazard rate, and the band's thickness indicates by how much. For example, at time zero, the treated judge hazard rate is 49% and the control judge hazard rate is 35%.

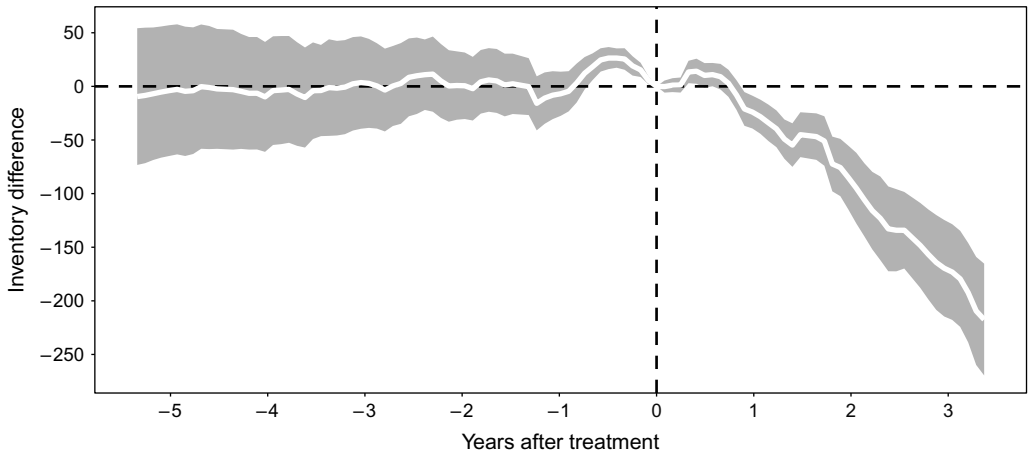
average treated judge adjudicated 1,807 hearings after the intervention (and before the May 31, 2014, data blackout) and thus finished $1{,}807 \cdot 0.0485 = 88$ more cases because of implementing case-level FIFO. After the intervention, the treated judges finished an excess of $88/1{,}247 = 0.07$ cases per day.

Figure 5 depicts our intervention's effect on hazard rates. To create this plot, we divided the sample into 100 time buckets, each comprising 1% of the treated judge hearings. For each bucket, we estimated the average treated and control judge hazard rates with the fraction of cases completed to hearings heard. The band's jagged bottom edge depicts the minimum of the treated and control time series, and the top edge depicts the maximum. The band is black when the treated series is larger, and gray when the control series is larger. The treated and control hazard rates mirror one another for five years before treatment—the band is thin with equal parts gray and black—but diverge soon thereafter—the band turns thick and black.

5.2. Inventory Decrease

Since hazard rates mediate our intervention's effect, we calculate the inventory decrease attributable to the hazard rate increases; doing so controls for unrelated arrival and service rate changes. Specifically, Figure 6 plots the reduction in case inventories over time attributable to the treated judges' abnormal hazard rates. To create this figure, we pair each treated judge with a counterfactual judge, who mirrors his counterpart in every way but one: the hazard rate. A counterfactual judge's hazard rate tracks the monthly average control judge hazard rate. For example, if a treated

Figure 6 Inventory Changes Attributable to Hazard Rate Differences



Notes. This plot depicts the difference between what the treated judges' inventory levels actually are and what they would have been had their hazard rates mirrored the control judges' hazard rates. We normalize the inventory difference to zero at the intervention date and depict 90% confidence intervals with gray bands.

judge finishes 35 cases out of 80 hearings in a month and the control judges collectively finish 450 cases out of 1,200 hearings, then the corresponding counterfactual judge finishes $(450/1{,}200) \cdot 80 = 30$ cases, and the treated judge's inventory falls by $35 - 30 = 5$ cases relative to the counterfactual. The graph depicts the mean deviation between the counterfactual judges' simulated inventories and the treated judges' true inventories. We normalize the difference to zero at the intervention date and bootstrap for 90% confidence intervals.

The graph falls sharply at the intervention date, diverging at an average rate of 64 cases a year. At horizon's end, the mean inventory difference has grown to 217 cases.[5] This inventory-reduction estimate is larger than that in Section 5.1 because it does not account for fixed effects, and it compares treated and control hazard rates month by month, rather than subsample by subsample.

5.3. Flow Time Decrease

Case flow times are too long to measure without censoring bias—the median case finishes after 1.78 years, 19% the length of our sample. Accordingly, we measure flow times via hearing age: the time between the file date and the first hearing and the time between subsequent hearings. Chopping the data set more finely in this manner enables us to salvage more of it; even if a case's conclusion is censored, its first few hearings still yield noteworthy timestamps. To formalize our flow time measure, consider a case filed on day t_0 with H hearings, in which the judge holds hearing $h \in \{1, \ldots, H\}$ on day t_h. The case's flow time decomposes into a sum of hearing ages: $t_H - t_0 = \sum_{h=1}^{H} a_h$, where $a_h = t_h - t_{h-1}$ is the age of hearing h (measured in days). So the expected case flow time equals the expected hearing age multiplied by an average of 2.3 hearings per case.[6]

We establish that our intervention decreased the treated judges' flow times with difference-in-difference regressions similar to those in Section 5.1. The only difference is that the dependent variable changes to hearing age a_h.

Table 4 presents the regression coefficients. The control hearing flow times average 264 days before intervention and $264 + 71 = 336$ days after; the treated hearing flow times average $264 + 69 = 333$ days before intervention and $264 + 71 + 69 - 48 = 356$ after. The *Post* · *Treated* interaction term suggests that adopting case-level FIFO decreased the average treated judges' hearing flow time by 48 days and case flow time by $48.1 \cdot 2.3 = 111$ days (12%). Note, this is actually a lower bound on the steady-state flow time decrease because transitioning to the efficient regime took time (see Figure 6).

Table 4 Flow Time Difference-in-Difference Regression Coefficients

	Without fixed effect		With fixed effect	
	Without controls	With controls	Without controls	With controls
Intercept	264.22	448.60	—	—
	(2.82)	(30.25)	—	—
Post	71.48	66.00	—	—
	(5.19)	(5.24)	—	—
Treated	68.57	66.49	—	—
	(9.42)	(9.76)	—	—
Post · *Treated*	−48.14	−46.06	−46.68	−42.55
	(13.65)	(13.84)	(6.98)	(7.43)

Notes. This table presents OLS regression coefficients. Each observation corresponds to a hearing. The dependent variable is flow time age a_h, the number of days between a case's current hearing and its previous hearing (or file date if it is the first hearing). The independent variables, controls, and fixed effects are as described in Table 3.

5.4. Appeals Rate Decrease

About a year after our intervention, the treated judges reported a serendipitous side effect: they forgot fewer case facts under case-level FIFO because of the reduced time between hearings. They speculated that better remembering the cases led to fairer rulings. Accordingly, we test whether switching to case-level FIFO improved the ruling quality with difference-in-difference regressions. We use the independent variables outlined in Section 5.1 but change the dependent variable to $100s_h$, where s_h is a dummy variable indicating that the case was appealed to the Supreme Court of Cassation. The expected value of this dependent variable is the case appeals rate in percent. Unjust rulings should be more frequently appealed (Coviello et al. 2015), so the treatment should decrease the dependent variable.[7]

Table 5 provides the regression coefficients. After our intervention, the rate at which the treated judges' rulings were appealed dropped by 3.8% relative to the control; this differential is sizable, as only 8.7% of treated cases are appealed. Thus, we find evidence that decreasing the time between hearings improved judicial outcomes. This increase in ruling quality was

[5] Since the hazard rate increase is temporary (our intervention cannot influence long-run throughput rates), the inventory deviation must eventually level off. We do not observe this plateauing because the effect is slow to materialize, because of the long flow times.

[6] Hearing flow times are also censored near the end of our horizon; to avoid censoring bias we remove hearings that arrive in the last year of our sample horizon. Because hearing flow times rarely exceed a year, only 0.5% of our remaining hearing flow times are censored. (This fraction is the same in our treated and control subsamples.) We also remove hearings that arrive in a judge's first and last years, to focus on steady-state performance.

[7] We observe *Appealed* for cases that finished before February 1, 2012, so we truncate our sample accordingly.

Table 5 Quality of Rulings: Difference-in-Difference Regression Coefficients

	Without fixed effect		With fixed effect	
	Without controls	With controls	Without controls	With controls
Intercept	7.78	−7.64	—	—
	(0.34)	(3.72)	—	—
Post	−6.27	−5.89	—	—
	(0.32)	(0.34)	—	—
Treated	3.62	3.49	—	—
	(0.68)	(0.60)	—	—
Post · Treated	−3.78	−2.46	−2.41	−1.54
	(0.65)	(0.64)	(0.67)	(0.64)

Notes. This table presents OLS regression coefficients. Each observation corresponds to a completed case (so *Post* now indicates that the case was completed after our intervention). The dependent variable is *Appealed*, a dummy variable indicating that the ruling was appealed to the Italian supreme court, multiplied by 100 (to express the rate of apeals as a percentage). The fixed effects, controls, and standard errors are as described in Table 3.

an unintended consequence; we did not anticipate that adopting case-level FIFO would reduce forgetting until after conducting the study. This fortuitous finding highlights the importance of measurement when recommending operational changes.

6. Robustness Checks

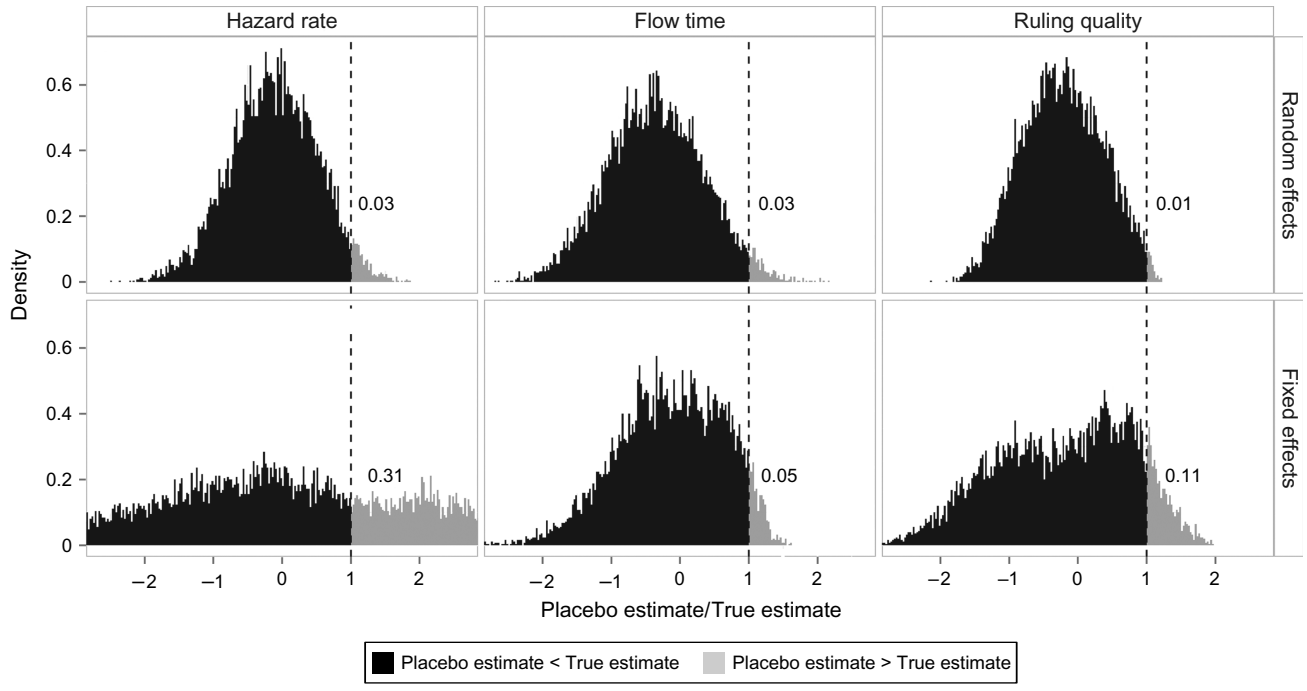
6.1. Placebo Test

Our data set is large—185,000 hearings spanning eight years—but our cross-section of treated judges is small—just six individuals. (It is not easy to compel judges to change their behavior.) Thus, when calculating standard errors, we rely heavily on temporal variation; specifically, when modeling a judge's behavior, we permit general autocorrelation within months but no autocorrelation across months. Thus, persistent temporal shocks may bias our standard errors, leading to spurious results.

To determine whether our difference-in-difference results are artifacts of a small treatment group, we conduct a placebo test. Specifically, we create 10,000 new samples by randomly assigning judges to *Treated* and *Control* groups; the case compositions, intervention date, and proportion of treated hearings remain fixed. For each sample, we run the control-free hazard rate, flow time, and ruling quality difference-in-difference regressions from Tables 3–5.

Figure 7 plots histograms of the *Post · Treated* coefficients, where the real estimates are normalized to one. Our true random effects estimates stand out relative to the simulations: out of 10,000 simulations, only 14

Figure 7 Placebo Test



Notes. This plot explores the robustness of our difference-in-difference estimates in light of our small sample of treated judges. Specifically, we consider the controls-free *Post · Treated* regression coefficients in Table 3, 4, and 5. We run equivalent difference-in-difference regressions for 10,000 synthetic data sets. We construct the synthetic data sets by randomly assigning judges to *Treated* and *Control* groups while fixing the fraction of treated judge hearings. We then plot the distribution of the ratio of the synthetic estimates to our actual estimates and report the fraction of synthetic estimates that exceed our actual estimates.

are stronger in both hazard and flow time, and only one is stronger in hazard rate, flow time, and ruling quality. And our true fixed effects estimates, although weaker, are also noteworthy: out of 10,000 simulations, only 368 are stronger in both hazard and flow time, and only 19 are stronger in hazard rate, flow time, and ruling quality. These results suggest our findings are not artifacts of a small treatment group.

6.2. Parallel Trends

For our difference-in-difference estimates to be valid, the treated and control hazard rates must exhibit parallel trend lines before intervention (Angrist and Pischke 2009, p. 230); the control sample would be a poor benchmark if it did not track the treated sample, preintervention. We test the parallel trends hypothesis by regressing our preintervention dependent variables on (i) the constant 1, (ii) *Time*, the number of centuries after the intervention (we use this timescale to scale up the regression coefficients), (iii) *Treated*, and (iv) the product of *Time* and *Treated*.

Table 6 tabulates the regression coefficients. The *Time* · *Treated* coefficient is statistically insignificant for hazard rate, flow time, and ruling quality dependent variables. Thus, we fail to reject the hypothesis that the treated and control hazard rates follow the same trend lines before our intervention.

Table 6 Preintervention Temporal Trend Regression Coefficients

	Hazard rate	Flow time	Ruling quality
Intercept	0.43	335.68	0.03
	(0.01)	(9.94)	(0.01)
Time	1.86	2,797.29	−1.89
	(0.30)	(281.26)	(0.52)
Treated	0.02	74.71	0.00
	(0.02)	(16.55)	(0.03)
Time · *Treated*	1.26	416.62	−1.53
	(0.69)	(618.99)	(1.19)

Notes. This table presents OLS regression coefficients. Each observation corresponds to a preintervention hearing. There are three dependent variables: (i) case completion dummy variable c_h, from the hazard rate regressions in Table 3, (ii) hearing age a_h, from the flow time regressions in Table 4, and supreme court appeal dummy variable s_h, from the ruling quality regressions in Table 5. The independent variables are *Time*, the number of centuries after the intervention, and *Treated* a treated judge dummy variable. We report robust month-judge block bootstrap standard errors in parenthesis.

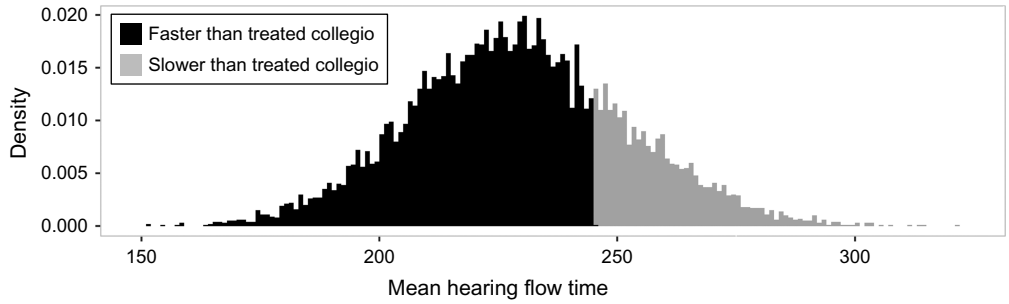
6.3. Self-Selection

Our treated subsample should be a fairly representative cross section of the court because: (i) the judges did not elect to participate in the study; instead, they were cajoled to join by the president of the *collegio*; (ii) all of the judges in the *collegio* agreed to the president's request; and (iii) *collegio* assignments are arbitrary, depending primarily on the court's availability at time of hire. Nevertheless, since the assignment of judges to treated and control subsamples was not strictly random, our estimates may suffer a self-selection bias: the treated judges might be more eager for improvement.

First, we test whether our treated judges had aberrant preintervention flow times. We compare our treated *collegio* to 10,000 randomly drawn six-judge groupings. We show in Figure 8 that our treated *collegio* is in the slowest 23% of six-judge combinations: slow, but not abnormally so.

Second, we test for a confounding effort effect with judicial activity difference-in-difference regressions. Our dependent variable is the number of days since the presiding judge last held a hearing (of any case): longer interhearing times indicate less active schedules and hence lower effort levels. Table 7 presents the regression coefficients. We find no evidence of a confounding motivation effect—and hence no evidence of a self-selection bias—because our intervention is not positively correlated with judicial activity. In fact, we

Figure 8 Preintervention Flow Time Distribution



Notes. This plot explores the robustness of our difference-in-difference estimates in light of our small sample of treated judges. Specifically, we consider the controls-free *Post* · *Treated* regression coefficients in Table 3, 4, and 5. We run equivalent difference-in-difference regressions for 10,000 synthetic data sets. We construct the synthetic data sets by randomly assigning judges to "treated" and "control" groups while fixing the fraction of treated judge hearings. We then plot the distribution of the ratio of the synthetic estimates to our actual estimates and report the fraction of synthetic estimates that exceed our actual estimates.

Table 7 Effort Levels: Difference-in-Difference Regression Coefficients

	Random effect		Fixed effect	
	Without controls	With controls	Without controls	With controls
Intercept	0.697	1.127	—	—
	(0.023)	(0.171)	—	—
Post	−0.217	−0.225	—	—
	(0.033)	(0.034)	—	—
Treated	−0.158	−0.154	—	—
	(0.046)	(0.047)	—	—
Post · Treated	0.133	0.120	0.104	0.127
	(0.068)	(0.069)	(0.081)	(0.082)

Notes. This table presents OLS regression coefficients. Each observation corresponds to a hearing. The dependent variable is the number of days since the presiding judge last held a hearing. The specifications with fixed effects include month and judge dummy variables; the specifications with controls include case type, plaintiff type, defendant type, and party count dummies. We report robust month-judge block bootstrap standard errors in parentheses.

find the opposite: the treated judges worked less frequently, relative to the control judges, after the intervention; specifically, the expected time a treated judge needed to execute a hearing increased by 0.13 days.

6.4. Hawthorn Effect

The Hawthorn effect could have biased our results: simply tracking the treated judges' performance could have increased their efficiency. But we find this unlikely because (i) the judges are not accountable to us (or to anyone, really); (ii) Figures 5 and 6 demonstrate that the effect held into the fourth year of implementation; and (iii) Table 7 demonstrates that the treated judges worked relatively less after the intervention.

7. Conclusion

This work identifies a new setting for operations management in the judiciary. Specifically, we study the Italian judiciary. This environment is ripe for operations research—it is critical, complex, and wasteful. But these courts are inefficient for a reason: They are hamstrung by byzantine procedural rules, an adversarial climate, bureaucratic inertia, and political strife. Moreover, judges, in our experience, are not predisposed to consult operations researchers. They are lawyers, not engineers, and they seek just rulings, not economical rulings. So it is not clear, a priori, that this institution would respond to operational prescriptions. Thus, this work serves as a proof of concept for operations management in the judiciary, aimed at judges and operations researchers alike. It shows that a simple insight from a multiarmed bandit model can decrease case flow times in the Roman Labor Court of Appeals by 111 days (12%), demonstrating that Italy's judicial gridlock is not entirely intractable. Using Coviello et al. (2015, p. 940) assessment that delaying an Italian labor case by a day decreases social welfare by €62, we estimate that our intervention increased social welfare by €105,690 per judge per year.

Acknowledgments

The authors thank Thomas Bray, Antonio Moreno, and Nicos Savva for their helpful suggestions; Amelia Torrice, Margherita Leone, and the other judges of the experimental court of Roma for changing their scheduling practices; and Vincenzo Ballarano and Antonio Simeone for data extraction.

References

Angrist JD, Pischke JS (2009) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton, NJ).

ANSA (2015) Giustizia: Ferie magistrati, Commissione Csm 'restano 45.' Accessed January 27, 2015, http://www.ansa.it/sito/notizie/politica/2015/01/27/giustizia-ferie-magistrati-commissione-csm-restano-45_94dff906-e4ac-42aa-b446-313c6a881cb7.html.

Aral S, Brynjolfsson E, Van Alstyne M, Van Alstyne M (2012) Information, technology, and information worker productivity. *Inform. Systems Res.* 23(3):849–867.

Arlotto A, Chick SE, Gans N (2014) Optimal hiring and retention policies for heterogeneous workers who learn. *Management Sci.* 60(1):110–129.

Batt RJ, Terwiesch C (2012) Doctors under load: An empirical study of service time as a function of census. Working paper, University of Pennsylvania, Philadelphia.

Berkowitz J, Kilian L (2000) Recent developments in bootstrapping time series. *Econometric Rev.* 19(1):1–48.

Caro F, Gallien J (2007) Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* 53(2):276–292.

CEPEJ (2006) Compendium of best practices on time management of judicial proceedings. Technical report, European Commission for the Efficiency of Justice, Strasbourg, France.

CEPEJ (2014) Report on European judicial systems—Edition 2014 (2012 data): Efficiency and quality of justice 2014. Technical report, European Commission for the Efficiency of Justice, Strasbourg, France.

Chirico A (2013) Le toghe e i 51 giorni di ferie: confessioni di un magistrato. *Panorama* (July 29), http://www.panorama.it/news/politica/ferie-magistrati-cappello/.

Coviello D, Ichino A, Persico N (2012) Time allocation and task juggling (preliminary draft). 104(2):1–26.

Coviello D, Ichino A, Persico N (2014) Time allocation and task juggling. *Amer. Econom. Rev.* 104(2):609–623.

Coviello D, Ichino A, Persico N (2015) The inefficiency of worker time use. *J. Eur. Econom. Assoc.* 13(5):906–947.

Davidson R, Mackinnon JG (2004) *Econometric Theory and Methods*, Vol. 21 (Oxford University Press, New York).

Esposito G, Lanau MS, Pompe S (2014) Judicial system reform in italy—A key to growth. Working paper, International Monetary Fund, Washington, DC.

Freeman M, Savva N, Scholtes S (2016) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Sci.*, ePub ahead of print September 2, http://dx.doi.org/10.1287/mnsc.2016.2512.

Gittins J, Glazebrook K, Weber R (2011) *Multi-Armed Bandit Allocation Indices*, 2nd ed. (John Wiley & Sons, New York).

Hardle W, Horowitz J, Kreiss JP (2003) Bootstrap methods for time series. *Internat. Statist. Rev.* 71(2):435–459.

Horowitz JL (2001) The bootstrap. Heckman JJ, Leamer E, eds. *Handbook Econometrics*, Vol. 5 (Elsevier Science, Amsterdam), 3159–3228.

Kc DS (2014) Does multitasking improve performance? Evidence from the emergency department. *Manufacturing Service Oper. Management* 16(2):168–183.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.

Mersereau AJ, Rusmevichientong P, Tsitsiklis JN (2009) A structured multiarmed bandit problem and the greedy policy. *IEEE Trans. Automatic Control* 54(12):2787–2802.

Niño-Mora J (2012) Towards minimum loss job routing to parallel heterogeneous multiserver queues via index policies. *Eur. J. Oper. Res.* 220(3):705–715.

Organisation for Economic Co-operation and Development (OECD) (2013) What makes civil justice effective? Technical Report 18, OECD Economics Department Policy Notes, Paris.

Pinedo M (2012) *Scheduling: Theory, Algorithms, and Systems* (Springer, New York).

Powell A, Savin S, Savva N (2012) Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing Service Oper. Management* 14(4): 512–528.

Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Sci.* 58(6): 1141–1159.

Stock JH, Yogo M (2005) Testing for weak instruments in linear IV regression. Andrews DWK, Stock JH, eds. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (Cambridge University Press, New York), 80–108.

Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Sci.* 60(6):1574–1593.

Unita' per la Costituzione (2015) Internal memo. Accessed November 21, 2015, http://www.unicost.eu/news/2015/11/relazione-roberto-carrelli-palombi-21_11_2015.aspx.

Wang L, Gurvich I, Van Mieghem JA, O'Leary KJ (2015) Task switching and productivity in collaborative work: A field study of hospitalists. Working paper, Northwestern University, Evanston, IL.

World Bank Group (2014) Doing business report: Enforcing contracts. Accessed January 1, 2015, http://www.doingbusiness.org/data/exploretopics/enforcing-contracts.