Service Operations for Justice-On-Time: A Data-Driven Queueing Approach

Nitin Bakshi
University of Utah, USA, nitin.bakshi@eccles.utah.edu

Jeunghyun Kim
Korea University Business School, South Korea, jeunghyunkim@korea.ac.kr

Ramandeep S. Randhawa
University of Southern California, USA, rrandhaw@marshall.usc.edu

Problem definition: Limited resources in the judicial system can lead to costly delays, stunted economic development, and even failure to deliver justice. Using the Supreme Court of India as an exemplar for such resource-constrained settings, we apply ideas from service operations to study delay. Specifically, court dynamics constitute a *case-management queue*, whereby each case may experience multiple service encounters spread across time, but all are necessarily with the same server. Our goal is to elucidate the drivers of congestion, focusing on metrics such as the expected case-disposition time (delay) and expected number of cases awaiting adjudication (pendency), and leverage this understanding to recommend operational interventions.

Methodology/results: We employ *data-driven calibrated simulations* to model the analytically intractable case-management queue. The life cycle of a case comprises two stages: pre-admission (before determining its merit for detailed hearings) and post-admission. Our methodology allows us to capture the queueing dynamics in which the judges are shared resources across the two stages. It also permits modeling of holiday capacity, which is flexibly tailored to address any surplus work that spills over from the regular year. We find that the second stage of this judicial queue is overloaded, but holiday capacity creates a perception of stability by steadying performance metrics.

Managerial implications: The sources of inefficiency that drive congestion include a misalignment between scheduling guidelines and judicial capacity, coupled with the requirement to schedule hearings in advance. Together, these factors inhibit utilization of shared capacity across the two-stage judicial queue. We demonstrate how interventions that account for these inefficiencies can successfully tackle judicial delay. In particular, scheduling to improve the allocation of time across pre- and post-admission cases can cut down the expected delay by as much as 65%.

Key words: judicial delay | case-management queues | data-driven simulation

All authors contributed equally to the paper.

1. Introduction

India has a population of about 1.38 billion people (The World Bank 2020a). The current backlog of cases in the Indian judicial system stands at nearly 40.6 million. About a quarter of these cases have been pending for more than five years (The Government of India 2021). Such judicial delays have adverse economic and social impact. For instance, in terms of "enforcing contracts" India is ranked 163 out of 190 nations (The World Bank 2020b), and about 70% of India's prison population comprises pre-trial detainees (World Prison Brief 2020).

Despite the workload, only about $21,000$ judge positions have been sanctioned to address the litigation that arises (Financial Times 2016). This translates to less than 16 judges per million people in contrast to the more aspirational figure of 50 judges per million (The Supreme Court of India 2012). Moreover, about $25\% - 33\%$ of India's sanctioned judicial strength remains unfilled (Singhvi 2017). While adding judges is one possible response to the situation, the judiciary itself has proposed multiple ways of tackling the problem of judicial delays, including improving process/procedural efficiency and using techniques for workload management (e.g., alternative dispute resolution mechanisms) (The Supreme Court of India 2018). The multiple options notwithstanding, progress in addressing judicial delays has been painfully slow. We believe that this is because the drivers of congestion in the judicial system are not well understood.

Queueing analysis, the workhorse model in service operations, offers an appropriate means to study congestion. The operational dynamics of the judicial system correspond to that of a case-management queue: a case that arrives typically has multiple hearings (service episodes) spread across time, but all of them with the same server (panel of judges) to whom the case is assigned – similar to a patient–doctor relationship in a medical setting. However, the case-management queue is known to be analytically intractable (Campello et al. 2017). In this paper, we therefore develop a *queueing-theoretic* simulation framework which we calibrate with data to first uncover the drivers of congestion, and then, to evaluate the relative effectiveness of interventions for managing delay.

We apply our framework and demonstrate its utility in the context of the Supreme Court of India (SCI). The reason to do so is twofold. First, the quality of data in the SCI is quite good relative to other lower courts in India (The Supreme Court of India 2012). Second, the SCI functions as a final court of appeals, and 80–90% of its workload comprises of appeals of decisions made by the state high courts (Robinson 2013a). Therefore, the SCI displays characteristics (high congestion levels) similar to other appeals courts even in developed nations (Bray et al. 2016, Green and Yoon 2017). Due to the co-existence of severe congestion and high-quality data, the SCI provides an ideal testbed for our framework.

1.1. Judicial Delay and Operations Management

Justice delayed is justice denied, is an age-old adage that is accompanied by periodic acknowledgement of the severity of the problem, particularly in appeals courts (Carrington 1969, Meador 1974, Adler 2014). The remedies are broadly classified into two categories: boosting infrastructure, and eliminating process inefficiency (e.g., Castro and Guccio 2015). The experience in India has been qualitatively similar albeit distinct in practice and procedure (Law Commission of India 1987, Malimath 2003, Law Commission of India 2014, The Supreme Court of India 2018). It was recognized a while back that the tools and concepts from service operations (e.g., queueing theory and management science), in combination with data, have the potential to make a big dent in addressing this critical societal challenge (Blumstein and Larson 1969, Nagel et al. 1978). However, notable contributions in this regard have been scarce.

A few exceptions are Bray et al. (2016), Azaria et al. (2023), and Azaria et al. (2024). Bray et al. (2016) investigates improved scheduling of cases within the Italian labor court of appeals. Specifically, the paper studies when the scheduling logic of *oldest-case-first* may perform better than *oldest-hearing-first* from the perspective of a judge who maximizes their long-term case disposal rate. Building on the theory of constraints, Azaria et al. (2023) develops an intervention in the form of a case-scheduling policy, and provides model-free evidence to support the intervention's effectiveness in the Jerusalem District Court in Israel. Azaria et al. (2024) conducts a more fine-grain statistical analysis in collaboration with the Jerusalem District Court to support their proposed case-scheduling policy. In this paper, our focus is on studying the benefits of operational interventions that have already been identified as promising directions by the SCI, which include optimizing capacity and process re-engineering (The Supreme Court of India 2018). However, any intervention that requires unpacking the legal aspects of decision making (e.g., workload management techniques such as alternative dispute resolution or mediated settlement) is beyond the scope of this paper because of the unavailability of data to support such analysis.

From a methodological perspective, we offer the analytical underpinnings for planning judicial capacity, and our approach is also related to a small but relatively influential stream of applied research that employs computational or simulation techniques to make progress in studying complex settings that are analytically intractable, e.g., Jordan and Graves (1995), Huchzermeier and Cohen (1996), Bakshi et al. (2011), Shumsky et al. (2021), Li et al. (2024).

1.2. Significance of Contribution and Methodology

As noted above, our work exemplifies how the rich body of knowledge within service operations can be innovatively adapted to generate transformational value in under-explored application areas such as the judiciary. Our main finding is that a direct consequence of the policies followed by the

SCI is that it operates like a two-stage queue in which the second stage is overloaded. Thus far, this realization has eluded practitioners and academics alike. We believe a key reason for this is the lack of insights into the dynamics of a two-stage case-management queue with shared server capacity (across the two stages) and the complex role stochasticity plays in impacting delay. Another reason for the inconspicuousness of the overload here is that by asking judges to work during the holidays, the SCI is able to flexibly deploy additional capacity to work off excess backlogs, thereby achieving stable performance metrics (delay and backlogs) across time, as shared in the court's annual reports (The Supreme Court of India 2019).

In terms of methodology, our model is calibrated with novel data, and has also benefited from interviews with the judges of the SCI, practicing senior lawyers at the supreme court, registrars at the court who are in-charge of recording new cases, and non-governmental organizations that promote judicial access and accountability.

We recommend analytically-derived guidelines for sharing judicial capacity across the two stages of the case-management queue such that congestion can be addressed in a cost-effective manner. The dynamics in a case-management queue with shared resources bear parallels to tandem queuing systems due to the two stages, yet are distinct due to the sharing of the resource (server) across stages. In particular, by appropriately partitioning the shared judicial capacity across stages, it is possible to better manage overall congestion, as well as the relative congestion across stages in the judicial queue, without having to increase total capacity. Our insights are useful not only for the SCI, but also for congested appeals courts across the world.

Finally, as explained before, we have steered clear of interventions that require tinkering with the quality and legal aspects of judicial decision making. But future research that looks into these questions, coupled with the insights from our work, has the potential to radically advance the field of judicial operations. For instance, the SCI conducts joint hearings for cases that may be filed by different parties (litigants) but share the same legal principles — such cases are referred to as connected matters. It could be rewarding to study how such joint hearings affect the quality of decision making versus delay, more so because there is an opportunity to establish connections with some insightful research that has already been conducted in the analogous context of shared-medical-appointments for healthcare delivery (Ramdas and Swaminathan 2021, Buell et al. 2024).

2. The Supreme Court of India (SCI)

The highest court in India's judicial hierarchy is its Supreme Court. It was established in 1950 with eight justices. Currently, it has a strength of 34 sitting judges, including the Chief Justice of India.[1] The SCI has original jurisdiction on matters such as protecting human rights; appellate

[1] The sanctioned strength of judges was increased to 34 in 2019. However, the actual strength can differ from the sanctioned strength depending on the timing of retirements and fresh appointments.

jurisdiction over decisions of lower courts that are appealed; and advisory jurisdiction on matters referred to it by the President of India (The Supreme Court of India 2023b). Appeals dwarf all other forms of workload, and this makes the SCI distinct from other institutions such as the U.S. Supreme Court, which restricts itself to constitutional matters and questions of extraordinary legal importance. Moreover, the nine justices of the U.S. court decide all matters together, or *en banc*. In contrast, to deal with its immense workload, the SCI adjudicates matters mainly through two-judge benches (Robinson 2013a). Each judge is meant to hear matters on every working weekday for at least 4.5 hours; more if there are pressing matters to be heard (The Supreme Court of India 2017b).

Overview of judicial operations at the SCI

The life cycle of a case in the SCI is essentially driven by multiple hearings with the same judicial bench that it gets assigned to upon arrival. The operational details about how the SCI functions are laid out in the official practice and procedure handbook (The Supreme Court of India 2017b). We provide a high level overview here.

Every day, a few hundred cases are filed, and they go through an administrative review process at the registry. The details of any new case, or *fresh matter*, are first recorded by the registrar. This requires complete and accurate documentation, hence, hearings practically cannot be scheduled on the same day that the case is registered. Thus, all cases that are registered on a particular day, are available for a potential hearing at the beginning of the next day, effectively *batching* their arrival. To ensure equitable workload distribution, a case is generally routed to the bench with the shortest queue, unless the CJI intervenes explicitly. Occasionally, a new case is connected to another case involving similar legal considerations, which has already been assigned to a bench. In that event, the new case is assigned to the same bench as the connected case, and they are heard together. Hearings at the SCI can result in different outcomes:

- Disposal: the judges arrive at their final decision after a hearing, and the case is disposed from the SCI
- Adjournment: a hearing is rescheduled because some elements are not ready for the hearing (for example, the litigants, lawyers or judges are unavailable for the hearing)
- Regular hearing: a follow-on hearing is required because the current hearing is not conclusive, and the case needs more judicial time before a decision can be reached
- Admission: the life cycle of a case potentially spans two sequential stages: the *pre-admission* stage and the *post-admission* stage. In the former stage, the SCI determines if the case even merits detailed scrutiny. Admission is that hearing outcome whereby a case transitions to the

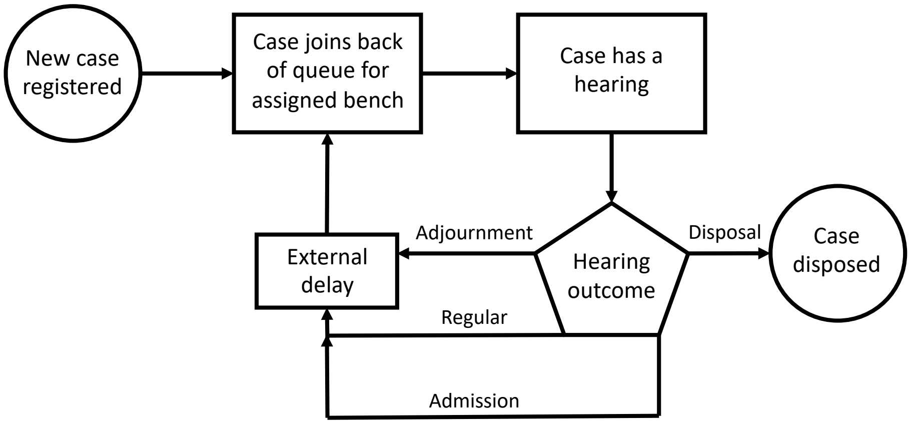

Figure 1 **Life cycle of a case at the SCI.**

latter (post-admission) stage for further hearings until a final decision on the case is reached, otherwise the case is disposed.[2]

After a regular hearing, an adjournment, or an admission, the case is sent to the back of the queue to await its next hearing (i.e., oldest-hearing-first). A peculiarity of the case-management queue at the SCI is that there are interludes between two consecutive hearings of a case, typically of the order of four to six weeks, such that irrespective of the congestion in the system, the case is unavailable to be scheduled for its next hearing during this time. We refer to such an interlude as *external delay*, and it is a judicially sanctioned window of time meant for completing essential tasks related to the legal aspects of the case, e.g., procurement of certain documents as evidence, or response to allegations, seeking consent of interested parties, etc.[3] We illustrate the life cycle of a case in Figure 1.

For scheduling of cases, the SCI organizes a work week into two buckets. On Mondays and Fridays, only pre-admission cases are scheduled to be heard. On the other hand, on Tuesdays, Wednesdays, and Thursdays, a mix of pre- and post-admission cases are scheduled for hearings.[4] Invariably, the court continues hearing matters during the vacation period (summer holidays). The schedule during the vacation period can differ from the standard schedule of a regular work week.

3. Model and Data

We capture the salient aspects of the SCI's workflow in our model, as described next. Specifically, we model the operational dynamics at the SCI as a *case-management queue* such that a case can have multiple hearings across time but always with the same bench (server).

[2] Based on our data, the percentage of cases admitted is 9.5%.

[3] Refer to Legros et al. (2020) for another example when external delay plays an important role in congested systems.

[4] This detail about scheduling is implicit in Figure 1, but is made explicit in Figure 2 together with the accompanying discussion in Section 3.3.

We simulate the performance of this judicial queue, and then use it to study congestion at the SCI. Calibrated simulation is the appropriate tool for analysis here because even the simplest variant of a case-management queue is known to be analytically intractable (Campello et al. 2017). Our setting is more complex than the canonical queueing settings for multiple reasons, including but not limited to the fact that the inter-arrival time and the service time are not exponential, and further, the server capacity is shared across two stages in the judicial queue.

We calibrate our model using "case status" data downloaded from the SCI website (The Supreme Court of India 2023b). We downloaded the details for cases that arrived between 2009 and 2016. Table 1 summarizes the arrival information between 2009 and 2016 based on the downloaded data.[5] Our data also tells us when a case was registered, the date for each hearing that it had, and the court's written orders associated with each hearing. The foundational elements for our model of the case-management queue are described in the next five subsections. The pseudo-code of our simulation that encompasses the elements described below can be found in Online Appendix A.

Table 1 Arrival of new cases to the SCI

Year	New cases registered
2009	39,811
2010	38,625
2011	38,032
2012	39,831
2013	39,834
2014	40,298
2015	39,130
2016	40,135
Total	315,696

3.1. The arrival process

We use the arrival information for cases from 2009 to 2016 for our simulation. Specifically, we simulate the arrival process to the SCI by maintaining the time and sequence of arriving cases as in the raw data.[6] To simulate the bench assignment for a case, we first search the list of outstanding cases in the simulated SCI to check for the existence of a case connected (legally) to the new case. If there is such a case then the new case is assigned to the same bench and heard together with this connected case. Information about connected cases is available in our data and we use it to assign new cases accordingly. For fresh cases without connection to existing cases, we assign them on the basis of the shortest queue-size when they arrive.

[5] We do not use the arrival information documented in the annual reports of the SCI; those numbers are tailored to the idiosyncratic accounting practices of the SCI which can lead to double counting of certain types of cases, and also suffer from other shortcomings that have been documented before (Robinson 2013a).

[6] For longer simulation runs that go beyond the time frame of the data, we re-initialize the arrivals from 2009, as many times as needed.

3.2. The service process

Work done to dispose off a case is referred to as service. It can be spread across multiple service encounters or hearings. Cases are heard by an assigned bench (panel of two judges) until the matter is resolved. We model 14 such homogeneous benches since the strength of the Supreme Court was 28 when we initiated this project in mid-2017.[7] Each bench maintains a dedicated queue of cases that are assigned to it.

A critical requirement for simulating the service process is the empirical distribution of time spent on an individual hearing. The details available on the case-status page of the SCI's website do not include information about the time spent on individual hearings in a case. Hence, we collected actual hearing times shown in real time by the online display-board on the court's website. We collected the data for a period of seven months, from November 12, 2018 to May 10, 2019. This resulted in $15,725$ observations. Those observations that spanned the lunch hour were appropriately corrected by deducting the extra hour from the hearing time.

The duration of hearings can vary considerably: from short hearings that last a few minutes to longer deliberations that run into hours. In the end, the empirical distribution yields a mean of 6.81 minutes and a standard deviation of 16.25 minutes. For our simulations, we sample the hearing time from this empirical distribution.[8]

3.3. Two-stage queue, hearing outcomes, and advance scheduling

Recall that the life cycle of a case potentially spans two sequential stages: the *pre-admission* stage and the *post-admission* stage. In the first (pre-admission) stage, the SCI determines if the case even merits detailed scrutiny. If the determination is that it does, only then is the case admitted to the second (post-admission) stage for further hearings until a final decision on the case is reached, otherwise the case is disposed. Thus, the post-admission stage has three types of outcome possible (*disposal*, *regular hearing*, *adjournment*), while the pre-admission stage has one additional outcome that is essentially whether the case moves on to the second stage (*admission*). Note that any case can only have one disposal, either from the pre-admission queue or from the post-admission queue.

An interesting and crucial aspect of SCI operations is that server capacity (judicial bench) is a shared resource across the two stages of the judicial queue. The SCI has issued guidelines for how judges will split their time across the pre- and post-admission stages (The Supreme Court of India 2017b). Specifically, on Mondays and Fridays, each bench aspires to hear 45 fresh matters

[7] In practice, some judges may also participate in specialist benches due to their expertise (e.g., income tax or environment related). The Chief Justice or registrar may assign cases to these specialist benches. However, such assignments are the exception not the norm; hence, we do not model specialist benches.

[8] In Online Appendix C.1, we investigate the robustness of our analysis with respect to the homogeneous hearing time assumption.

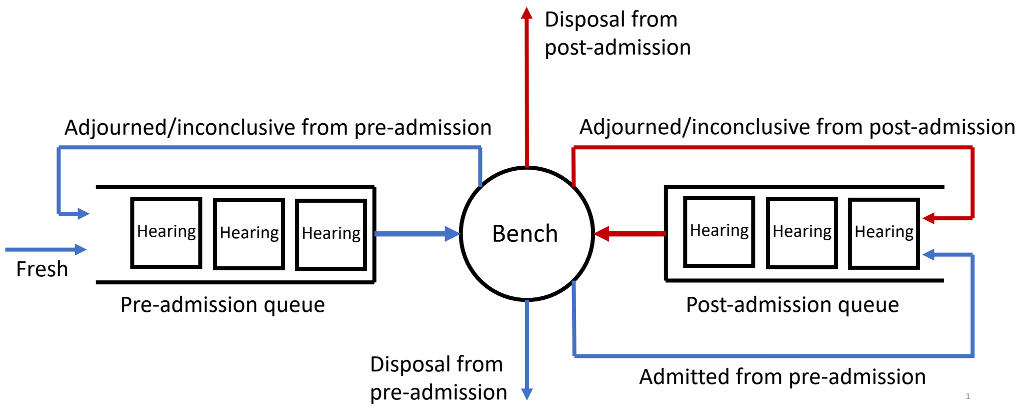

Figure 2 **Two-stage judicial queue at the SCI.**

(first hearing) and 15 more advanced pre-admission cases, also called *after-notice* matters. On the remaining days, the intent is for each bench to hear daily about 10 after-notice cases and 10 post-admission cases. For these days, the scheduling guidelines explicitly allow for the discretion of the Chief Justice of India (CJI). Indeed, prior research suggests that the number of after-notice cases heard in a day have been significantly higher than the recommended number 10 (the exact number has varied depending on the CJI); while the number of post-admission hearings were between 10 and 20 (Sindhu and Narayan 2018). Consistent with modeling a unique hearing-time distribution, we set the total number of cases to be scheduled on each day to be the same, i.e., 60, and we use the number of post-admission hearings to calibrate our simulation. In our simulation, we capture the guidelines from the SCI by modeling each bench as a multi-stage, single-server queue illustrated in Figure 2.

Importantly, the cases to be heard on a particular day are finalized a few days in advance and shared with the concerned parties through a document known as the *cause list*. This allows the concerned parties to prepare, and if required, make travel plans for the hearing. Consequently, it is infeasible to make changes to the cause list (scheduled cases) on short notice.

Estimating the decision probabilities: The outcome of a hearing is not recorded explicitly in the data. Fortunately, we have access to the written court orders issued at the end of every hearing. We conducted keyword-based text analysis of the court order files to classify the outcomes.

This classification is quite challenging because the orders can vary in their descriptiveness and also have significant variety in how various outcomes were described. Using a combination of key words and length of the order as the means for classification, we created a decision tree that is depicted in Figure 3. We also sought expert inputs from practicing supreme court lawyers to refine and corroborate our classification logic.

To assess the accuracy of our decision tree, we randomly sampled 500 orders and applied our classifier to them. We also manually read each order to ascertain the outcome directly. Our algorithm correctly classified 462 hearings, which implies an accuracy of 92.4% with its 95% confidence interval being [90.1%, 94.7%]. We document the confusion matrix associated with the accuracy testing in Online Appendix B.

Based on this classification algorithm, we determined the fraction of hearings that resulted in a *disposal*, *adjournment*, *regular hearing*, or *admission* to the second stage. We interpret these fractions as the probability of the corresponding outcome for a hearing. Due to the potential for these probabilities to vary over time, we tracked their evolution across years and then settled on two sets of numbers that correspond to different regimes of decision making: one set corresponds to hearings before 2015 (Table 2), and the other for hearings conducted after 2015 (Table 3).[9] In our simulation, we implement hearing outcomes by random sampling in an *i.i.d.* fashion based on the information in Table 2 or 3.

Table 2 Decision probabilities before 2015

Stage	Disposal	Adjournment	Regular	Admission
Pre-admission	0.38	0.37	0.24	0.01
Post-admission	0.09	0.69	0.22	NA

Table 3 Decision probabilities after 2015

Stage	Disposal	Adjournment	Regular	Admission
Pre-admission	0.36	0.37	0.19	0.08
Post-admission	0.24	0.57	0.19	NA

3.4. External delay

We extract the required statistical information about external delay from the written court orders issued at the end of a hearing. This information is conveyed in the order files in the form of a statement, e.g., “list after 4 weeks.” The summary statistics for external delay are presented in Table 4. In our simulation, at the end of a regular, admission, or adjourned hearing, we sample from the empirical distribution of external delay. The case is then unavailable for its next hearing during the realized value of the external delay.

[9] We have also considered, and ruled out, the possibility that the admission probability is a function of prevalent congestion level, as measured through backlogs at the time the admission decision is made. Refer to Online Appendix D for details.

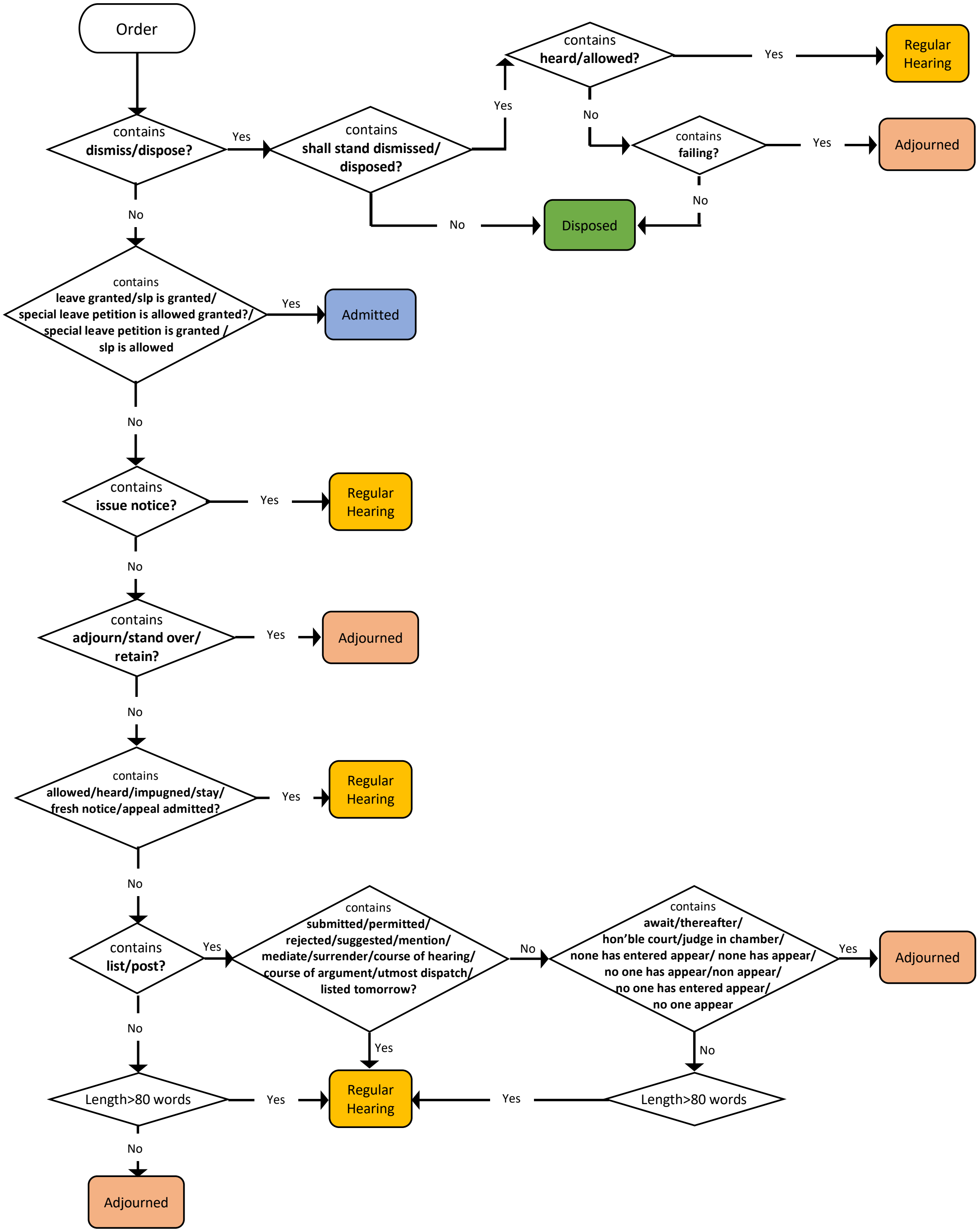

Figure 3 Decision tree to classify hearing as disposed, admitted, adjourned or regular.

Table 4 Summary Statistics for External Delay

Stage	Mean	Standard Deviation
Pre-admission	3.55 wks	4.42 wks
Post-admission	3.40 wks	5.02 wks

3.5. Holiday capacity

Annual reports of the SCI reveal that invariably the court continues hearing matters during the vacation periods with the *explicit goal to prevent backlogs exceeding an acceptable level.*[10] In a year, the number of hearings during holidays have typically been in the range of $1,000$ to $2,000$ (The Supreme Court of India 2015, 2016, 2017a).

In our model, "holiday" actually refers to weekdays during the 7-weeks of SCI summer vacation. There is no formal documentation of how this capacity is organized at the SCI, except what we can infer from the summary descriptions in the annual reports.

The way we have modeled holiday operations is as follows. On each day during this "holiday" duration, we first compare the number of outstanding cases in the system with a pre-determined threshold. If the number of outstanding cases exceeds this threshold, we trigger each bench to process post-admission cases (which we identify as the driver of congestion) for the duration of a regular workday. We use the triggering threshold to calibrate our simulation.

Another aspect to consider is the cost of using holiday capacity. Should we be thinking in terms of hourly remuneration of the judges? Perhaps it is more appropriate to think in terms of how holiday work substitutes the time of judges away from careful writing of judgments, and keeping up with latest developments in the legal space (Pandey 2023). Already, due to paucity of time, the judges of the SCI have sacrificed writing comprehensive judgments (and setting precedent for future cases with analogous legal content) in favor of hearing more cases: The SCI issues only about $1,000$ judgments per year (Chandra et al. 2018). This has contributed to the criticism that the SCI is *polyvocal*, i.e., it does not systematically build on precedent but different benches take different positions on similar questions of law (Robinson 2013b, 2014).

It is noteworthy that organizing judicial work during vacations is an intervention by the CJI to tackle congestion, which takes the form of vacation or holiday benches. As we shall see, small as it is, holiday capacity nonetheless plays an important role in the operational performance of the judicial queue.

[10] Such goals, as espoused by the CJI, are often described in the annual reports of the SCI along with summarizing various interventions to counter congestion, including constituting holiday benches for the summer and winter breaks. For instance, "All out efforts are being made to reduce the pendency of cases to about $50,000$ in the near future," (The Supreme Court of India 2015), and, "A sustained effort to clear the backlog has resulted in reducing the pendency of this court below the $60,000$ mark by the end of the year 2015," (The Supreme Court of India 2016). The exact number is not so important because of the SCI's different accounting practices mentioned earlier (Robinson 2013a).

4. Results and Analysis

When analyzing queueing systems, the expected delay (disposition time) experienced by the cases is a natural metric to consider. Another metric of interest is the expected backlog of pending cases (pendency). These two metrics of delay and pendency are related through *Little's law*: expected backlogs = expected arrival rate × expected delay. Formally, one is interested in the long-run average measure of these quantities, assuming the system is stable and ergodic (Wolff 1989). Intuitively, if cases arrive at a higher rate than can be processed, the system would be unstable and the backlog of cases would continue increasing over time. This notion is captured via the (theoretical) *utilization* of the system, denoted by ρ, which measures the ratio of system demand to the system capacity. A situation of $\rho \geq 1$ implies instability and would be associated with very high delays and pendency, which theoretically would continue to increase with increase in time. If utilization is less than 100%, i.e. $\rho < 1$, then ρ also equals the fraction of time the system capacity is busy in processing work. Further, in this case, one can analyze the system's steady state characteristics. Unfortunately, the analysis of case-management queues is very complex and analytical characterization of the expected delay is not possible in generality. In what follows next, we describe the results of our approach based on calibrated-simulations to address these challenges.

4.1. Strategy for baseline analysis

Our first step is to calibrate the simulations with data along the lines of the description in §3. Through this exercise, we want to set the baseline by capturing the first-order effects in SCI operations; specifically, we aim to match the first moment of simulated delay to the expected delay observed in our raw data. For the latter calculation, we worked with all disposed cases in our dataset (besides the cases mentioned in Table 1, we also had access to partial data for the years 2008 and 2017), to arrive at an expected delay (disposition time) of 275.6 days based on $319,471$ observations.

Besides monitoring the expected delay (disposition time), the SCI tracks instances of *excessive delay* to individual cases (The Supreme Court of India 2012). In other words, the SCI cares about the thickness in the right-tail of the delay distribution. In fact, as per the Malimath Committee's recommendations, any matter that is delayed for more than two years should be treated as arrears[11] and prioritized (Malimath 2003). Thus, we say that a case has experienced excessive delay if it has been delayed beyond two years The raw data reveals that about 13% of the SCI cases experience excessive delay.[12]

Other than workload-driven congestion, a factor that exacerbates delay is the notion of "stay orders" whereby the SCI imposes a temporary injunction on court proceedings, resulting in delays

[11] The state of being behind in the discharge of obligations.

[12] The Normal distribution has a Kurtosis equal to 3. The delay distribution's Kurtosis > 3, indicating a right skew.

that often exceed a decade (Deshpande 1986). Thus, in terms of the time-frames involved, stay orders are well-beyond the scope of our dataset even though they contribute to the right tail of the delay distribution. In combination with other unobservables such as the discretion of the CJI in scheduling cases (on Tuesdays, Wednesdays, Thursdays; refer §3.3), this inhibits our ability to exactly match higher-order effects on delay. Hence, beyond matching the first-moment of the delay distribution, we aim to only qualitatively parallel the thickness in its right tail, as embodied in excessive delay. This approach allows us to focus on capturing the underlying *operating regime* of the SCI.

Although benchmarking our baseline simulations against the empirical delay distribution is sensible, we still need to know the exact number of hours each day that the court operates. SCI's procedure handbook does prescribe a daily duration of at least 4.5 hours for judges to conduct hearings; this excludes time spent offline in reviewing case briefs and writing judgments (The Supreme Court of India 2017b). But a number of hearings (e.g., urgent bail applications) may be conducted within the chambers of the judges and not in the courtrooms, and are thus not accessible to public nor is the duration observable on the display-board (The Supreme Court of India 2015, p. 75). As a result, we do not have complete visibility into the realized duration of a workday. Our conversations with judges of the SCI confirmed that their workday is closer to about 5 hours.

To conduct simulations, we are still missing some critical information: 1) the split between after-notice and post-admission matters on Tuesday/Wednesday/Thursday, while respecting the contextual observations described in §3.3; and 2) the backlog level targeted through holiday capacity. The standard statistical approach to deal with missing or unobservable data is to use imputation (Efron 1994). In our context, we impute the value of the two missing parameters to achieve the following three targets: to match the expected disposition time in the raw data; to replicate the reliance on holiday capacity; and to mimic excessive delay.

4.2. Simulating the baseline and insights

4.2.1. Calibration: Through our calibrated simulations, we achieved a match along the three dimensions described previously. Specifically, by scheduling 15 post-admission matters followed by 45 after-notice matters on Tuesdays to Thursdays, and using holiday capacity (if required) to trim backlogs below $32,000$ cases, we obtain an expected disposition time of 275 days (in all simulations, half-widths of the 95% confidence intervals for expected durations are less than half a day).[13, 14] The fraction of cases that experience excessive delay is 17.6%.[15]. Holiday capacity is used for an average of $1,657$ hearings per year.

[13] Note that on Mondays and Fridays, as the cases are all pre-admission matters, we randomize between fresh matters and after-notice.

[14] All confidence intervals are reported in Online Appendix F.

[15] Kurtosis is > 3, consistent with the notion of a right skew in the delay distribution.

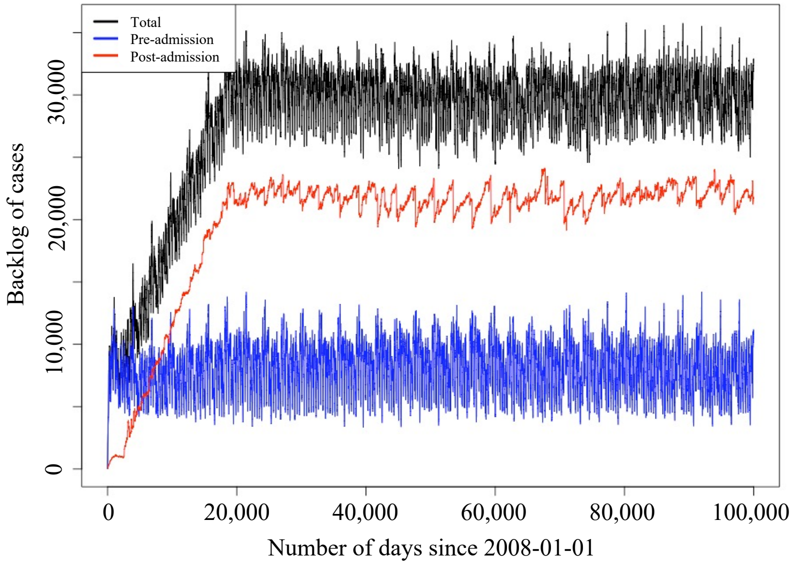

Figure 4 Simulated backlog of cases: with holiday capacity.

4.2.2. Insights: Overloaded second stage and use of holiday capacity: Having matched the key operational metrics of the SCI to gain confidence in our baseline, we pursue a closer examination of the simulated outcome. It is interesting to note that the expected delay in the pre-admission stage of the judicial queue is only 75 days compared to an expected delay of 1144 days in the post-admission stage. Further, we note that after a case has been admitted to the second stage, the time until the first hearing is on average about 1124 days or nearly 3 years. [16] Thus, it appears that it is the second stage of the judicial queue that is disproportionately congested. This is consistent with Sindhu and Narayan (2018), which, based on interviews, states that there is insufficient time to hear regular hearing cases. We next ponder the nature of this congestion, and also its key drivers.

It is possible to determine the utilization of each stage in the judicial queue by direct examination of the simulation output. It turns out that the utilization of the pre-admission stage is 82.9% and that of the post-admission stage is 101.1%, which confirms that the second stage is overloaded. Yet, we observe stable overall metrics such as a finite expected delay of 275 days and expected backlogs are at $27,731$ cases. (Refer to Figures 4 and 5 and for a visual rendition of the time-evolution of backlogs, with and without holiday capacity; only the former is stable.) The reason for this stability is that the SCI flexibly deploys holiday capacity to work through any excess congestion to achieve the targeted backlog number.

[16] This is consistent with what we learnt from our interviews with lawyers and judges of the SCI. FAQ's on the SCI's website further corroborate the finding that an admitted case has to wait multiple years before it can be heard (The Supreme Court of India 2023a, p. 2).

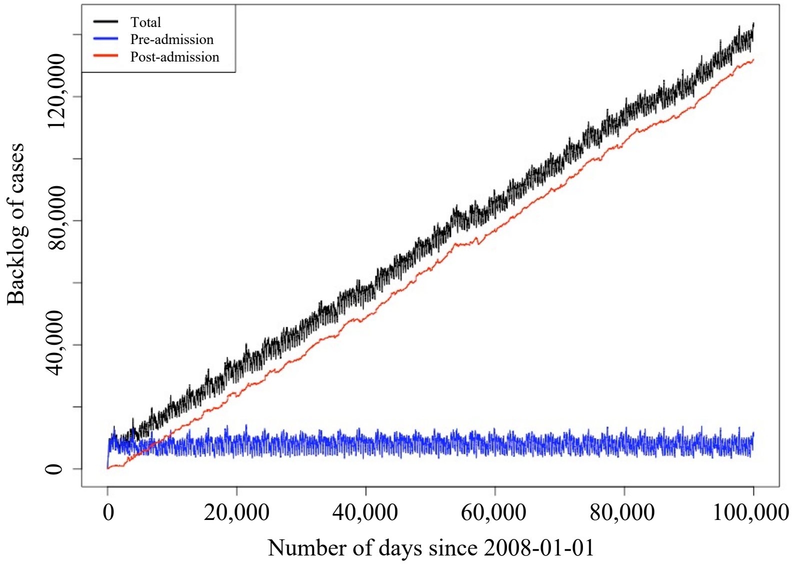

Figure 5 **Simulated backlog of cases: without holiday capacity.**

Therefore, we refer to the operating regime of the SCI as being *quasi-stable*. It is the use of holiday capacity that restores stability at the expense of judicial "overtime". In §3.5, we discussed that the direct and indirect cost of this overtime can be very significant. It is therefore reasonable to conclude that use of holiday capacity is an undesirable way to counter congestion, especially if there are alternatives. At this point, it is unclear how the quasi stable regime may respond to classical interventions such as an increase in service capacity. We explore this next.

4.2.3. Insights: Impact of boosting capacity by adding judges: An obvious approach to reduce delay is to appoint more judges to the SCI. We have modeled only two-judge benches in our case-management queue, hence, the smallest increase in judicial capacity would be the addition of a 15th bench, i.e., increasing the number of judges from 28 to 30. Using exactly the same parameter values as imputed in our baseline simulation, we summarize our findings in Table 5.

By adding a bench (7% increase in capacity), the expected disposition time dropped by 8% to 253 days, while excessive delay remained the same (up to some simulation variation). Although the holiday hearings decreased in number, holiday capacity is invoked yet again because the post-admission queue is overloaded, i.e., the judicial queue continues to be quasi-stable. Thus, the operating regime of the SCI is relatively insensitive to small changes in capacity.

Adding enough additional judges will undoubtedly resolve congestion, but it is not straightforward to achieve. Adding judges not only has financial implications related to salary and infrastructure, but its feasibility is also linked to the availability of suitably qualified judges in large numbers.

Table 5 Impact of Adding Judges

	14 benches	15 benches
Expected delay	275 days	253 days
Excessive delay	17.6%	17.9%
Holiday hearings	1657 per year	84 per year

Furthermore, due to its wide-ranging implications, changing the capacity at the SCI requires legislation. Therefore, we continue our search for alternatives. To identify these alternatives, we must uncover the reasons for operational inefficiency in the judicial queue.

4.2.4. Identifying sources of operational inefficiency: A peculiarity of the judicial queue is its two-stage nature with shared capacity. This gives rise to new considerations such as how to appropriately partition service capacity across the two stages. In particular, similar to tandem queues, there is potential to *starve* or *flood* the downstream queue (post-admission) depending on the throughput from the upstream queue (pre-admission). But different from tandem queues, the service capacity is shared across the two stages. Hence, by scheduling a certain fixed number of cases on Monday-Friday and Tuesday-Wednesday-Thursday, it is possible to mitigate *capacity imbalance* in a way that favorably influences congestion and performance in the judicial queue.

Moreover, scheduling of cases by itself does not entirely determine the available capacity for the pre- and post-admission stages. This is because the hearing times are random and if the scheduled hearings finish early (before the end of a workday) then there are no more hearings that can be conducted that day even though backlogs may be nonzero (recall discussion about advance scheduling in §3.3). We refer to this situation as *forced idleness.*[17] The opposite outcome occurs if hearings run over and certain cases cannot be reached on the scheduled day. Such cases have to be rescheduled for another day. Part-heard matters (incomplete hearings) are taken up for hearing the next day.

4.3. Identifying operational improvements

4.3.1. Value of rebalancing capacity: The recourse to counter the two sources of inefficiency—capacity imbalance and forced idleness—is to revisit the scheduling of cases across weekdays. To illustrate this point, we revert to our baseline capacity of 14 benches, but now we schedule 17 post-admission cases followed by 43 after-notice cases on Tuesday to Thursday (daily total is fixed at 60 hearings). By increasing the capacity allocated to the previously "overloaded" post-admission queue, we find that performance of the judicial queue improves dramatically, as

[17] Note that we have modeled the workday as being used exclusively to conduct hearings. No doubt, judges do a lot preparatory work beyond simply conducting hearings. But that occurs outside of the workday duration and we do not model that time. Hence, upon an early end to hearings, even if judges redirected their time towards tasks other than hearings, we treat that time as idle time from a queuing perspective.

reported in Table 6: the expected disposition time is now 96 days (a reduction of 65%), and excessive delay is only 0.1%. This is because the post-admission stage is not overloaded anymore, as is evident from the absence of holiday hearings.

We note that the significant improvement in performance metrics (the expected disposition time, the number of hearings taking place during summer, and the probability of experiencing excessive delay) does not hurt the pre-admission queue even though our intervention is about reallocating capacity towards the post-admission queue. While our intervention reduces the expected post-admission delay from 1144 days down to 115 days, the expected delay in the pre-admission queue remains at almost the same level (75 days) before and after the intervention.

To highlight the benefit of our intervention, we convert the elimination of holiday hearings to savings in judicial working time across the SCI. Given that average hearing time is 6.8 minutes, our intervention saves 187.8 hours a year, on average, of judicial working time. Additionally, elimination of holiday hearings would free up judges from preparing for these hearings outside their regular working hours. We believe that these savings are significant because of how time-constrained judges are; the extra time could be redirected towards systematic documentation of judgements that have the potential to set legal precedent.

The table also shows the downside of the intervention: it increases the number of cases that were scheduled for hearing on a given day but not reached because the workday ended before their turn.

Our intervention—changing the TWTh split from 15/45 to 17/43 and hearing post-admission cases before after-notice matters—increases the number of unreached post-admission cases because there is no guarantee that, on a given day, every scheduled case will be heard that day itself. By changing the split from 15/45 to 17/43, each bench schedules two more post-admission cases a day. However, we observe in our simulation that the workday for judges often ends before hearing them all. Hence, the number of unreached post-admission cases goes up as seen in Table 6.

Further, our intervention increases the number of unreached pre-admission (after-notice on TWTh) cases. This is because, on some days, the prevailing backlog of after-notice cases that are not under external delay is not high enough to schedule 43 such hearings. In fact, we observe that on average 29 after-notice hearings are scheduled per day per bench on TWTh. After implementing our intervention, which shifts capacity allocation towards post-admission cases, the backlog of after-notice cases increases, and hence, we are able to schedule more than 29 after-notice cases. This increases the probability of the workday ending before all scheduled after-notice matters are heard. Accordingly, the average number of unreached after-notice hearings per day increases.

An increase in the number of unreached cases can be inconvenient for the involved lawyers and litigants, particularly in the post-admission stage, because the parties have to prepare and travel (often across the country) for a scheduled hearing.

Table 6 Rebalancing Capacity between the stages improves performance considerably

	15/45 split (TWTh)	17/43 split (TWTh)
Expected delay	275 days	96 days
Excessive delay	17.6%	0.1%
Holiday hearings	1657 per year	0 per year
Forced idleness	14%	12%
Post-admission cases not reached (TWTh, all benches)	1.68 per day	2.43 per day
Pre-admission cases not reached (TWTh, all benches)	119 per day	135 per day

Altering the split of scheduled cases to favor post-admission cases has to be done carefully because it may have implications for system performance. In particular, increasing the scheduled number of post-admission cases may leave inadequate capacity for after-notice matters, thereby overloading the pre-admission queue instead.

Overall, though, it seems that rebalancing capacity is a promising intervention. Moreover, the correct partitioning of capacity, through scheduling of cases, ought to be tailored to the number of judicial benches, and even to the arrival rate of cases to the SCI. Presently, and as discussed previously, the procedure handbook of the SCI provides a static guideline for scheduling cases that is independent of the number of serving judges and the arrival rate of cases. While the guideline makes allowance for the CJI to exercise discretion, as the above analysis demonstrates, this is not a trivial task. Next, we provide an analytical characterization that offers the desired guidance.

Analytical sufficient conditions for stability: We have discussed that the performance metrics for the two-stage judicial queue with shared capacity are analytically inaccessible. This is true even for determining exact capacity utilization of the individual stages because the number of cases available to schedule on a given day is a function of the backlogs, which itself is intractable. Even without an analytical characterization for utilization, we know that the judicial queue will be stable if, for each category of cases (fresh matters, after-notice matters, post-admission case), there is adequate capacity available to address the work associated with that category.[18] We leverage this insight to derive sufficient conditions for the stability of the judicial queue, which in turn, ensures that congestion is contained.

Let n_w^f, n_w^a, n_w^p be, respectively, the daily number of fresh, after-notice, and post-admission cases that a bench intends to hear on weekday $w \in \{\mathrm{MF}, \mathrm{TWTh}\}$. We set $n_{\mathrm{MF}}^p = n_{\mathrm{TWTh}}^f = 0$ because the SCI does not hear post-admission cases on MF or fresh matters on TWTh. We have modeled a random ordering of scheduled cases on MF, but through our analysis, we imputed a strictly higher priority for post-admission cases on TWTh. Specifically, on MF, if both fresh and after-notice cases are available for the next hearing, a fresh matter is heard with probability $q_{\mathrm{MF}} := n_{\mathrm{MF}}^f / \left(n_{\mathrm{MF}}^f + n_{\mathrm{MF}}^a\right)$,

[18] We will utilize an approach that partitions capacity by category to derive the sufficient conditions for stability. Our calculation also accounts for forced idleness in measuring the minimum capacity available for each category of work.

while an after-notice hearing is chosen with probability $1 - q_{\mathrm{MF}}$. We use the following additional notation:

- λ: the arrival rate per year of fresh cases to the SCI, where we count connected matters as one unit;
- μ: the service rate of a bench measured in hearings per minute, so that $1/\mu$ is the average hearing time in minutes;
- p_i^j: the probability that the outcome of a hearing at stage-i is j, where i refers to the stages of *Pre-admission (Pre)* or *Post-admission (Post)* and j refers to the outcome of a hearing, *Disposal (Disp)*, *Adjournment (Adj)*, *Regular(Reg)*, or *Admission (Adm)*;
- d: the number of working days per year;
- $\mathcal{D}$: the workday duration in minutes;
- b : the number of benches.
- $\{H_w^j(i)\}_{i\geq 1}$: a sequence of i.i.d. random variables that follow the hearing-time distribution. In this notation, workday $w \in \{\mathrm{MF}, \mathrm{TWTh}\}$, and hearings are of type $j \in \{f, a, p\}$, which respectively stands for fresh matter, after-notice, and post-admission hearings. Note that $\mathbb{E}[H_w^j] = 1/\mu$, and we assume that $H_w^j \leq \mathcal{D}$.

a. Stability of the pre-admission queue. The arrivals to this queue have two sources: the fresh arrivals to SCI, and repeat hearings (after-notice) for pre-admission cases before their disposal or admission to the post-admission queue. We will ensure there is sufficient capacity to handle arrivals from each source.

We will partition the workday by q_{MF} and assume that $q_{MF}\mathcal{D}$ is the amount of workday available for fresh matters.[19] Then, the workload of fresh matters that a bench can handle in a day ($w =$ MF) can be lower bounded by the minimum of the total work associated with incoming fresh matters, $\sum_{i=1}^{n_{MF}^f} H_{\mathrm{MF}}^f(i)$,[20]and the time available with a bench to process the work, which is $q_{MF}\mathcal{D}$ less the amount of time the bench must spend on left-over work from the previous day (part-heard matter). The time spent on the part-heard matter is upper bounded by the length of one entire hearing, which we denote by H_0. This random variable, H_0, is i.i.d. to all other hearing time random variables, hence, we can lower bound the time available with a bench to process the incoming work

[19] It is possible that more time than $q_{\mathrm{MF}}\mathcal{D}$ is available to hear fresh matters if the backlog of after-notice matters approaches zero. But that does not pose a problem because we are seeking sufficient conditions for stability.

[20] Technically, this assumes there is a sufficient backlog of fresh matters that are available for scheduling (i.e., not under external delay) such that this backlog exceeds n_{MF}^f for each bench, else we would modify this statement taking a minimum with backlog of available cases. However, because our goal is to identify sufficient conditions for stability (which still allows substantial congestion), it suffices to consider situations with a large enough backlog of available cases. The same idea also applies in the calculation of the workloads, W_{MF}^a and W^p, later in this discussion.

by $q_{MF}\mathcal{D} - H_0$. This implies that W^f defined below serves as a lower bound on the total workload of fresh matters that a bench can process in a day:

$$W^f := \min\left\{ q_{MF}\mathcal{D} - H_0, \sum_{i=1}^{n^f_{MF}} H^f_{\mathrm{MF}}(i) \right\}. \tag{1}$$

Importantly, the definition of W^f accounts for the forced idleness stemming from the fact that a bench cannot hear more than n^f_{MF} cases, even if the workday is not fully utilized after hearing n^f_{MF} fresh matters. (The same logic applies to the random variables that represent the number of after-notice and post-admission hearings defined in the next paragraphs.) Because 40% of the working days are MF, the annual processing time available at the SCI to process fresh-matter hearings is:

$$\mathcal{F} := 40\% \times d \times b \times \mathbb{E}\left[W^f\right]. \tag{2}$$

Because λ/μ is the rate at which fresh matter work arrives to the system, adequate capacity is ensured by satisfying the following condition,

$$\lambda/\mu < \mathcal{F}. \tag{Condition 1}$$

Assuming (Condition 1) is satisfied, the processing of fresh matters gives rise to an arrival rate of $\left(p^{\mathrm{Adj.}}_{\mathrm{Pre}} + p^{\mathrm{Reg.}}_{\mathrm{Pre}}\right)\lambda$ for after-notice cases. While classified as after-notice, each case can potentially have multiple hearings, which inflates the annual arrival rate for after-notice hearings to:

$$\lambda^a := \left(p^{\mathrm{Adj}}_{\mathrm{Pre}} + p^{\mathrm{Reg}}_{\mathrm{Pre}}\right)\lambda / \left(p^{\mathrm{Disp}}_{\mathrm{Pre}} + p^{\mathrm{Adm}}_{\mathrm{Pre}}\right). \tag{3}$$

The amount of work associated with after-notice hearings on Mondays and Fridays that a bench handles is at least W^a_{MF}, where

$$W^a_{\mathrm{MF}} := \min\left\{ (1 - q_{\mathrm{MF}})\mathcal{D} - H_0, \sum_{i=1}^{n^a_{\mathrm{MF}}} H^a_{\mathrm{MF}}(i) \right\}. \tag{4}$$

On the other hand, for Tuesdays, Wednesdays, and Thursdays, a bench first hears post-admissions hearings, and is therefore capable of processing at least W^p amount of work corresponding to post-admission hearings, where

$$W^p := \min\left\{ \mathcal{D} - H_0, \sum_{i=1}^{n^p_{\mathrm{TWTh}}} H^p_{\mathrm{TWTh}}(i) \right\}. \tag{5}$$

Because the bench hears after-notice matters only after working through the scheduled post-admission cases, this gives us the following lower bound on the amount of daily work associated with after-notice hearings that a bench can handle on these days:

$$\begin{aligned} W^a_{\mathrm{TWTh}} &:= \left(\min\left\{ \mathcal{D} - H_0 - \sum_{i=1}^{n^p_{\mathrm{TWTh}}} H^p_{\mathrm{TWTh}}(i), \sum_{i=1}^{n^a_{\mathrm{TWTh}}} H^a_{\mathrm{TWTh}}(i) \right\} \right)^+ \\ &= \min\left\{ \mathcal{D} - H_0 - W^p, \sum_{i=1}^{n^a_{\mathrm{TWTh}}} H^a_{\mathrm{TWTh}}(i) \right\}. \end{aligned} \tag{6}$$

Therefore, noting that 60% of the working days are Tuesdays, Wednesdays, and Thursdays, and combining with (2), the annual capacity available at the SCI to process after-notice hearings is

$$\mathcal{A} := 40\% \times d \times b \times \mathbb{E}\left[W^a_{\mathrm{MF}}\right] + 60\% \times d \times b \times \mathbb{E}\left[W^a_{\mathrm{TWTh}}\right]. \tag{7}$$

Adequate capacity for after-notice matters is ensured by satisfying the following condition,

$$\lambda^a/\mu < \mathcal{A}. \tag{Condition 2}$$

b. Stability of the post-admission queue. Assuming (Condition 1) and (Condition 2) are satisfied, fresh matters reach the post-admission queue at the rate, $p^{\mathrm{Adm}}_{\mathrm{Pre}}\lambda$, while after-notice matters reach the post-admission queue at the rate, $p^{\mathrm{Adm}}_{\mathrm{Pre}}\lambda^a$. The combined rate, $p^{\mathrm{Adm}}_{\mathrm{Pre}}(\lambda + \lambda^a)$, is inflated due to repeat hearings to yield an arrival rate of hearings to the post-admission queue:

$$\lambda^p := p^{\mathrm{Adm}}_{\mathrm{Pre}}(\lambda + \lambda^a)/p^{\mathrm{Disp}}_{\mathrm{Post}}. \tag{8}$$

Noting that the post-admission cases are strictly prioritized on TWTh, the number of post-admission hearings that the SCI can annually process without using holiday capacity is,

$$\mathcal{P} := 60\% \times d \times b \times \mathbb{E}\left[W^p\right]. \tag{9}$$

Hence, for the post-admission queue to be stable without the use of holiday capacity, it is sufficient that,

$$\lambda^p/\mu < \mathcal{P}. \tag{Condition 3}$$

Thus, it follows directly that:

Theorem 1 *If the three conditions: $\lambda/\mu < \mathcal{F}$ (Condition 1), $\lambda^a/\mu < \mathcal{A}$ (Condition 2), and $\lambda^p/\mu < \mathcal{P}$ (Condition 3) are met, then the judicial queue is stable without requiring any holiday capacity.*

We now evaluate the three conditions in the context of the SCI. We compute $\mathbb{E}[W^f]$, $\mathbb{E}[W^a_{\mathrm{MF}}]$, $\mathbb{E}[W^a_{\mathrm{TWTh}}]$, and $\mathbb{E}[W^p]$ using the technique of bootstrapping (Efron and Tibshirani 1994). Setting the model primitives to be $n^f_{\mathrm{MF}} = 45$, $n^a_{\mathrm{MF}} = 15$, $n^a_{\mathrm{TWTh}} = 43$, and $n^p_{\mathrm{TWTh}} = 17$, we note that (Condition 1), (Condition 2), and (Condition 3) are satisfied. Consistent with Theorem 1, our simulations confirm that the judicial queue is stable (see Table 6).

We also note that using the baseline model primitives $n^f_{\mathrm{MF}} = 45$, $n^a_{\mathrm{MF}} = 15$, $n^a_{\mathrm{TWTh}} = 45$, and $n^p_{\mathrm{TWTh}} = 15$, we do find our conditions no longer hold, specifically (Condition 3) is violated. While this does not guarantee instability because our conditions only provide sufficient conditions for stability, our simulations illustrate that indeed the judicial queue is unstable for these parameters (see Figure 5). This gives us confidence that our sufficient conditions are not too conservative.

Using the theoretical analysis for capacity planning: We would like to point out that our theoretical analysis of stability, at a given capacity level and arrival rate, can also help determine the target number of judges (capacity) to ensure stability for a different rate of litigation (arrival rate of cases). Theorem 1 can still provide the conditions that guarantee stability, which can be complemented with the simulations to estimate the performance metrics associated with the recommended configuration.

We believe this could be a major improvement over the current approach for capacity planning. The report of the 245th Law Commission of India favors the "Rate of Disposal Method" (Law Commission of India 2014, p. 24).[21] It relies on a deterministic treatment to ensure that "… the number of disposals [of cases] *equals* the number of institutions in any one year." Although intended to ensure stability, the SCI's approach fails to account for the two-stage queuing dynamics of the judicial queue, as embodied in Theorem 1.

4.3.2. Value of limiting adjournments: India's Code of Civil Procedure, 1908, lays down the guidelines and recommended time-frames for the completion of various procedural steps. These guidelines are often flouted, thereby contributing to congestion. A particularly troublesome feature of the current equilibrium is the high number of adjournments granted in some cases. An amendment to the Code of Civil Procedure in 1999 required a cap of three adjournments in any suit. Subsequent decisions by the SCI have diluted the status of this amendment to a mere recommendation (Ranjan 2016). After all, capping adjournments can be a concern if the litigants have a genuine reason. Nevertheless, eminent jurists and policy makers continue to champion the merits of capping adjournments, and contrasting with the status in countries such as USA and UK, where the number of adjournments granted is far less.

[21] The report considers, but does not find suitable, another capacity planning framework that it refers to as "Time-Based Method," which is also used in the United States.

Table 7 Low Impact of Limiting Adjournments in Pre-Admission Stage

Reduction in Adjournment prob.	Expected Delay	Excessive Delay
0% (base case)	275.46 days	17.6%
5%	273.98 days	17.6%
10%	274.87 days	17.5%
15%	275.3 days	17.6%
20%	272.37 days	17.5%

Table 8 High Impact of Limiting Adjournments in Post-Admission Stage

Reduction in Adjournment prob.	Expected Delay	Excessive Delay
0% (base case)	275.46 days	17.6%
5%	92.44 days	0.1%
10%	87.4 days	0%
15%	84.74 days	0%
20%	82.84 days	0%

Our approach of using calibrated simulations allows us to run *what-if* analysis to determine the performance impact of limiting adjournments. Given that both sides of the debate on adjournments are hotly contested, it is not our intent to be prescriptive. Rather, our goal is to strengthen the ability of policymakers to weigh the *pros* and *cons* of various options. Specifically, we contrast the performance impact of reducing the adjournment probability at the pre-admission stage versus the post-admission stage. To do so, for the baseline case with 14 benches and a 15/45 split on TWTh, we reduce the adjournment probability by a certain amount and redistribute this weight among the other hearing outcomes (regular, disposal, etc.) in proportion to their respective weights.

The results of our analysis are summarized in Tables 7 and 8. The simple and impactful message from these results is that adjournment of cases in the post-admission stage is much more consequential for overall congestion than adjournment in the pre-admission stage. Moreover, most of the gains are accrued by reducing adjournment probability by just 5% (the expected delay goes down to 92.44 days), further reduction is not of much value. The reason, once again, pertains to the post-admission stage of the judicial queue being overloaded prior to the intervention.

Given that only a small reduction in post-admission adjournments is required to reduce delay substantially, there is reason to hope that this can be achieved through process re-engineering, without impacting decision making from a legal perspective. In particular, Galanter and Robinson (2013) note the phenomenon of “grand advocates”, whereby an elite group of lawyers practicing at the SCI (about 40 to 50 of them) control a bulk of the litigation work at the court. The authors further note that, “Judges give postponements [adjournments] to eminent lawyers, who are constantly juggling more courtroom obligations than they can possibly attend . . . There is a cartelization of litigation by the senior counsel. They take on more cases than they can deal with. This leads to delays in the courts because the seniors don’t make it for appearances.” These observations suggest that to the extent an updated scheduling logic can recognize and avoid conflicts for the grand advocates, the number of adjournments at the post-admission stage can be mitigated without adversely affecting any of the parties.

Thus, both rebalancing capacity and targeted reduction of adjournments are promising interventions. Inputs from policy makers will help determine the extent to which they need to be melded with practical constraints to strike the right balance.

5. Robustness

We have developed a data-driven simulation model based on queueing theory. As one may imagine, it is not possible to exactly replicate the functioning of a complex institution such as the SCI. Instead, our goal is to create a model that is tractable yet realistic enough to generate robust results. Below, we revisit two key assumptions and discuss why we believe our analysis and insights are robust.

Assumption 1: Ignoring history dependence

We have assumed that the outcome of a hearing for a case is independent of how long the case has been pending. This may not be entirely accurate because, for instance, the annual reports of the SCI describe ad-hoc initiatives of the court to expedite disposal of excessively delayed matters, e.g., (The Supreme Court of India 2015, p.72-77), (The Supreme Court of India 2019, p.83-84). It is infeasible to model such discretionary interventions with reasonable precision, and they are targeted at only a small proportion of cases. Moreover, to the extent that expediting leads to an early hearing, independent of the remaining work needed to resolve the case, and therefore does not indicate a change in the expected outcome of a hearing, we can invoke the principle of work-conservation in queueing systems.[22] Correspondingly, we would expect the first moments of our key metrics (e.g., average disposition time) to be robust to any such scheme for prioritizing among cases (Wolff 1970).

Assumption 2: Ignoring case heterogeneity

The SCI classifies cases by the nature of the matter under consideration, e.g., civil versus criminal matters. However, we have not tried to capture this scheme to segment the population of cases. Instead, we have treated the cases as homogeneous. This assumption simplifies our exposition and treatment while still being aligned with our high-level objective of tracking performance statistics at the level of the entire population (not by case type). All the same, we believe our insights are practical and robust to this assumption for the following reason. From a perspective of controlling congestion, the key consideration is to achieve stability, which we demonstrate can be achieved by

[22] A change in the legal outcome based on scheduling is unacceptable from the standpoint of *jurispudence*.

proper scheduling that appropriately allocates capacity across the pre- and post-admission stages. Therefore, the marginal benefit from leveraging case heterogeneity (e.g., by prioritizing "shorter" cases) will be small. (More discussion in Online Appendix E.) Further, it is unclear if consistently prioritizing certain categories of cases is an acceptable policy, as it leads to concerns about fairness (Sindhu and Narayan 2018). Perhaps for these reasons, the SCI has never adopted such a policy.

We have also conducted additional robustness checks that are reported in Online Appendix C. Specifically, we assumed that hearing times are identically distributed because our hearing-time dataset is not granular enough to differentiate further. However, in Online Appendix C.1, we demonstrate that our insights are not limited by this assumption. Finally, to increase the confidence in our baseline calibration, we report a rolling window cross-validation exercise in Online Appendix C.2.

6. Discussion

The discipline of operations management has long been viewed as a promising means to address judicial congestion. However, substantial contributions in this regard have been few and far between. By marrying the knowledge of queuing theory with data that became recently available, we believe our framework with calibrated simulations delivers on this promise. Specifically, by modeling judicial operations as a novel two-stage queue with shared capacity, we are able to uncover deep insights pertaining to optimal scheduling and capacity planning at the SCI. Moreover, our *what if* analysis reveals a disproportionate impact of controlling adjournments in the post-admission stage of the judicial process. These previously unknown results are data-driven and readily actionable, thus offering valuable guidance to policy makers in the judicial context.

The entire focus of this research has been on more efficiently organizing judicial operations so as to minimize congestion resulting from a given arrival rate. We have not attempted to address the question of whether the arrival rate can be controlled, for instance, by being more selective in the appeals which are allowed to reach the SCI. This question has been explicitly dealt with by Sindhu and Narayan (2018), who state the following with regard to controlling judicial congestion:

> Two broad approaches to solving this problem stand out. One is for the SCI to devise policies or guidelines that would discourage the institution of certain cases, which could include identifying kinds of cases that would not be heard under the SCI's Art. 136 jurisdiction [*appeals*], and making advocates accountable for continuously filing cases that do not fall within this jurisdiction. The second approach would be for the SCI to maintain a rate of disposal that is higher than its rate of institution, so that it can gradually reduce backlog, while also hearing the fresh cases instituted before it.

The authors go on to explain that the first approach is infeasible because the CJI cannot possibly devise a policy that would discourage the institution of cases, for the following three reasons:

1. Such a policy could reasonably be challenged on the grounds that it is beyond the ambit of the CJI's administrative power.
2. The SCI has modeled itself as a "people's court," meant to remedy all forms of injustice. Any broad policy that restricts appeals will be seen as a withdrawal of this protection.
3. Finally, ignoring these concerns about the court's social legitimacy, there are still structural impediments to overcome. Any policy that seeks to contain the number of appeals filed would have to be implemented uniformly across multiple benches, each of which makes decisions "independently". As such, the structure of the SCI makes it difficult to limit appeals in a consistent fashion.

Having noted the above opinion of legal academics, it is worth emphasizing that our framework is agnostic to the arrival rate. Using our *what if* approach, it is straightforward to evaluate the impact of curtailing the arrival rate and estimating the new performance metrics. The reason we do not report such analysis in this paper is that its practical relevance is not clear in light of the observations in Sindhu and Narayan (2018).

It is also useful to dwell on the research opportunities that open up based on our work. The analytical treatment of the two-stage case-management queue with shared capacity is challenging, but it is a worthy ambition, especially the possibility of applying simpler models (e.g., fluid models) in the overloaded and quasi-stable regimes. Similarly, as mentioned in §1, field studies on the impact of connected matters is a potentially rewarding direction to pursue. These questions are of general importance, transcending the boundaries of judicial operations.

We have also discussed that one of the inefficiencies resulting from the way operations are organized at the SCI, is the impact of unheard (not reached) matters on litigants and lawyers. Could this concern be alleviated by reserving a certain number of slots for hearings in which the concerned parties voluntarily sign-up, with the explicit awareness that these hearings will have the lowest priority in the schedule for that day? Framed differently, a field study could evaluate the promise in matching a hearing slot with parties with a higher tolerance for being rescheduled. Given the potential for wastage, it makes sense to consider as candidates for *voluntary* hearings, those with shorter and more predictable hearing times.

Although we have demonstrated the utility of our framework in the context of procedures followed at the SCI, it can be adapted and applied fruitfully to address the capacity planning needs of India's lower judiciary as well. Equally, we hope that our framework will help in combating congestion in courts in other parts of the world. Overall, we hope that our work will stimulate interest within the operations community to further study how to improve the efficiency of decision making in the justice delivery process.

Acknowledgement: This research was accomplished with crucial inputs from the judges of the Supreme Court of India, practising senior lawyers at the supreme court, registrars at the court who are in-charge of recording new cases, and non-governmental organizations that promote judicial access and accountability. The authors gratefully acknowledge helpful conversations with Justice Deepak Gupta and Justice Madan B. Lokur, former judges of the Supreme Court of India; Mr. Gaurav Pachnanda and Mr. Sahil Tagotra, senior advocates practicing in the Supreme Court of India; Justice Sanjay Parihar, former Registrar of the Supreme Court of India; Mr. Harish Narasappa and Mr. B.S. Surya Prakash from DAKSH, an organization that promotes judicial access and accountability in India. The authors are also grateful for the research assistance provided by Mr. Shagun Gupta, Mr. Arsh Sidana, Mr. Bhavya Kaushik, Mr. Archit Matta, and Mr. Damoun Rabie. This study is (partially) supported by Korea University Business School Research Grant.

References

Adler, Andrew L. 2014. Extended vacancies, crushing caseloads, and emergency panels in the federal courts of appeals. *J. App. Prac. & Process* **15** 163.

Azaria, Shany, Boaz Ronen, Noam Shamir. 2023. Justice in time: A theory of constraints approach. *Journal of Operations Management* **69**(7) 1202–1208.

Azaria, Shany, Boaz Ronen, Noam Shamir. 2024. Alleviating court congestion: The case of the jerusalem district court. *INFORMS Journal on Applied Analytics* **54**(3) 267–281.

Bakshi, Nitin, Stephen E Flynn, Noah Gans. 2011. Estimating the operational impact of container inspections at international ports. *Management Science* **57**(1) 1–20.

Blumstein, Alfred, Richard Larson. 1969. Models of a total criminal justice system. *Operations Research* **17**(2) 199–232.

Bray, Robert L, Decio Coviello, Andrea Ichino, Nicola Persico. 2016. Multitasking, multiarmed bandits, and the Italian judiciary. *Manufacturing & Service Operations Management* **18**(4) 545–558.

Buell, R., K. Ramdas, N. Sönmez, K. Srinivasan, R. Venkatesh. 2024. Shared service delivery can increase client engagement: A study of shared medical appointments. *Manufacturing & Service Operations Management* **26**(1) 154–166.

Campello, Fernanda, Armann Ingolfsson, Robert A Shumsky. 2017. Queueing models of case managers. *Management Science* **63**(3) 882–900.

Carrington, Paul D. 1969. Crowded dockets and the courts of appeals: the threat to the function of review and the national law. *Harvard Law Review* 542–617.

Castro, Massimo Finocchiaro, Calogero Guccio. 2015. Bottlenecks or inefficiency? an assessment of first instance italian courts' performance. *Review of Law & Economics* **11**(2) 317–354.

Chandra, A., W. Hubbard, S. Kalantry. 2018. The supreme court of india: An empirical overview of the institution. *U. Chicago, Working paper* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3154597.

Deshpande, V.S. 1986. Stay orders—abuse. *Journal of the Indian Law Institute* **28**(2) 141–168.

Efron, B. 1994. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* **89**(426) 463–475.

Efron, Bradley, Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Financial Times. 2016. India's top judge Thakur pleads for help with avalanche of cases. https://www.ft.com/content/788fd8f8-0aac-11e6-b0f1-61f222853ff3?mhq5j=e7.

Galanter, M., N. Robinson. 2013. India's grand advocates: a legal elite flourishing in the era of globalization. *International Journal of the Legal Profession* **20**(3) 241–265.

Green, Andrew, Albert H Yoon. 2017. Triaging the law: Developing the common law on the supreme court of india. *Journal of Empirical Legal Studies* **14**(4) 683–715.

Huchzermeier, Arnd, Morris A Cohen. 1996. Valuing operational flexibility under exchange rate risk. *Operations research* **44**(1) 100–113.

Jordan, William C, Stephen C Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management science* **41**(4) 577–594.

Law Commission of India. 1987. Manpower Planning in Judiciary: A Blueprint. http://lawcommissionofindia.nic.in/old_reports/rpt120.pdf.

Law Commission of India. 2014. Arrears and Backlog: Creating Additional Judicial (wo)manpower. http://lawcommissionofindia.nic.in/reports/Report_No.245.pdf.

Legros, B., O. Jouini, O. Z. Akşin, G. Koole. 2020. Front-office multitasking between service encounters and back-office tasks. *European Journal of Operational Research* **287**(3) 946–963.

Li, Bingxuan, Antonio Castellanos, Pengyi Shi, Amy Ward. 2024. Combining machine learning and queueing theory for data-driven incarceration-diversion program management. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38. 22920–22926.

Malimath, V.S. 2003. Committee on Reforms of Criminal Justice System, Government of India, Ministry of Home Affairs. https://indialawyers.files.wordpress.com/2009/12/criminal_justice_system.pdf.

Meador, Daniel J. 1974. Appellate courts: Staff and process in the crisis of volume. *St. Paul, MN: West Publishing Co* .

Nagel, Stuart, Marian Neef, Nancy Munshaw. 1978. Bringing management science to the courts to reduce delay. *Judicature* **62** 128.

Pandey, R. 2023. Is Vacation necessary for judges and Lawyers? https://www.thelawadvice.com/articles/is-vacation-necessary-for-judges-and-lawyers.

Ramdas, Kamalini, Soumya Swaminathan. 2021. Patients could share virtual medical appointments for better access to telemedicine. *Nature Medicine* **27**(1) 14–16.

Ranjan, Brajesh. 2016. What causes judicial delay? Judgments diluting timeframes in code of civil procedure worsen the problem of adjournments. https://timesofindia.indiatimes.com/blogs/toi-edit-page/what-causes-judicial-delay-judgments-diluting-timeframes-in-code-of-civil-procedure-worsen-the-problem-of-adjournments/.

Robinson, Nick. 2013a. A quantitative analysis of the indian supreme court's workload. *Journal of Empirical Legal Studies* **10**(3) 570–601.

Robinson, Nick. 2013b. Structure matters: The impact of court structure on the Indian and US Supreme Courts. *The American Journal of Comparative Law* **61**(1) 173–208.

Robinson, Nick. 2014. Judicial architecture and capacity https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2505523.

Shumsky, Robert A, Laurens Debo, Rebecca M Lebeaux, Quang P Nguyen, Anne G Hoen. 2021. Retail store customer flow and covid-19 transmission. *Proceedings of the National Academy of Sciences* **118**(11).

Sindhu, J., V. Narayan. 2018. Institution matters. *Law and Politics in Africa, Asia and Latin America* **51**(3) 290–331.

Singhvi, Abhishek. 2017. Open letter to incoming CJI: The most pressing priorities before India's judiciary, and how to address them.

The Government of India. 2021. National Judicial Data Grid. https://njdg.ecourts.gov.in/njdgnew/index.php.

The Supreme Court of India. 2012. National Court Management System – Policy and Action Plan. https://main.sci.gov.in/pdf/NCMSP/ncmspap.pdf.

The Supreme Court of India. 2015. Annual Report 2014. https://main.sci.gov.in/publication.

The Supreme Court of India. 2016. Annual Report 2015-16. https://main.sci.gov.in/publication.

The Supreme Court of India. 2017a. Annual Report 2016-17. https://main.sci.gov.in/publication.

The Supreme Court of India. 2017b. Handbook On Practice and Procedure and Office Procedure. https://main.sci.gov.in/practice-and-procedure.

The Supreme Court of India. 2018. Conference Proceedings of National Initiative to Reduce Pendency and Delay in Judicial System. https://districts.ecourts.gov.in/sites/default/files/Proceeding%20Book%20Supreme%20Court.pdf.

The Supreme Court of India. 2019. Indian Judiciary: Annual Report 2018-19. https://main.sci.gov.in/publication.

The Supreme Court of India. 2023a. Frequently asked questions for advocates/litigants. https://main.sci.gov.in/php/FAQ/5_6246991526434439182.pdf.

The Supreme Court of India. 2023b. Website of the Supreme Court of India. https://main.sci.gov.in/.

The World Bank. 2020a. Population, total - India. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=IN.

The World Bank. 2020b. Ranking economies by ease of doing business. https://www.doingbusiness.org/en/data/exploreeconomies/india.

Wolff, Ronald W. 1970. Work-conserving priorities. *Journal of Applied Probability* **7**(2) 327–337.

Wolff, Ronald W. 1989. *Stochastic modeling and the theory of queues*. Pearson College Division.

World Prison Brief. 2020. World Prison Brief data. https://www.prisonstudies.org/country/india.

ONLINE APPENDIX.

Service Operations for Justice-On-Time: A Data-Driven Queueing Approach

Nitin Bakshi
University of Utah, USA, nitin.bakshi@eccles.utah.edu

Jeunghyun Kim
Korea University Business School, South Korea, jeunghyunkim@korea.ac.kr

Ramandeep S. Randhawa
University of Southern California, USA, rrandhaw@marshall.usc.edu

Appendix A: Pseudo-code of the simulation

In this section, we provide a pseudo-code for our simulation. To make it self-contained, we first list and define terms used in the pseudo-code.

- Case type (fresh matter, after-notice, and post-admission): Each case potentially goes through two sequential stages, the pre-admission stage and the post-admission stage. The former stage is associated with two case types, fresh matter and after-notice. A fresh matter is a newly registered case waiting for its first hearing at the SCI. A more advanced case in the pre-admission stage is called after-notice. The case's type is converted to post-admission only when the SCI determines that the case merits detailed scrutiny. In our simulation, cases are tracked and backlogged in their type-specific queues.
- Cause list: It is a list of cases scheduled for hearing on a day.
- External delay: It is a judicially sanctioned window of time between two consecutive hearings of a case. External delay is meant for completing essential tasks related to the legal aspects of the case and thereby the case is not available for hearing until external delay is exhausted.

Step 1. Arrival and bench assignment

- Task 1: Register a batch of daily arrivals (cases) to the system.
- Task 2: For each new case, check for existence of a connected case in the system backlogs. If there is such a case, tag the new case to the connected case and have them heard together. Otherwise, assign the new arrival to the back of the fresh-matter queue of a bench with the smallest backlogs. Once all cases in the batch are routed, go to Step 2.
- Data source: For this step, we use a sequence of daily case arrivals from 2009 and 2016.
- Assumption: We assume that 14 benches in our simulation are homogeneous.

Step 2. Check date

- Task: If the date belongs to a summer holiday, go to Step 10. Otherwise, go to Step 3.

Step 3. Check the day of the week

- Task: If the day is Monday or Friday, go to Step 4. If the day is Tuesday, Wednesday, or Thursday, go to Step 5. Otherwise, go to Step 9.

Step 4. Monday and Friday cause list generation

- Task 1: From the front of the fresh matter queue, add 45 cases (whose external delays are exhausted) to the cause list.
- Task 2: From the front of the after-notice queue, add 15 cases (whose external delays are exhausted) to the cause list.
- Task 3: Go to Step 6.
- Assumption: Position in the queue is preserved in the cause list.

Step 5. Tueday, Wednesday, and Thursday cause list generation

- Task 1: From the front of the post-admission queue, add 15 cases (whose external delays are exhausted) to the cause list.
- Task 2: From the front of the after-notice queue, add 45 cases (whose external delays are exhausted) to the cause list.
- Task 3: Go to Step 7.
- Assumption: Position in the queue is preserved in the cause list.

Step 6. Monday and Friday case selection for hearing

- Task 1: Check whether the cause list is empty or the remaining daily capacity of the bench is 0. If it is, go to Step 9.
- Task 2: If the cause list is not empty, observe composition of the list.
- Task 3: If cases of both types are in the list, choose the first fresh-matter case in the list with 75% chance and choose the first after-notice case in the list with 25% chance. Otherwise, choose the first case of existing type in the cause list.
- Task 4: Go to Step 8.

Step 7. Tuesday, Wednesday, and Thursday case selection for hearing

- Task 1: Check whether the cause list is empty or the remaining daily capacity of the bench is 0. If either is true, go to Step 9.
- Task 2: If the cause list is not empty, observe whether there are post-admission cases are in the list.
- Task 3: If so, choose the first post-admission case in the list. If not, choose the first after-notice case in the list.
- Task 4: Go to Step 8.

Step 8. Hearing

- Task 1: If the case is part-heard from the previous day, use the carried over hearing time as a hearing time for this part-heard case. Otherwise, sample a hearing time for the chosen case.
- Task 2: Compare the hearing time with the remaining daily capacity of the bench.

- Task 3: If the remaining capacity of the day exceeds the hearing time, go to Step 8.1. Otherwise, go to Step 8.2.
- Data source: Hearing times were collected from the SCI's online display board in a period spanning from November 12, 2018 to May 10, 2019
- Assumptions: We assume that hearing times across types are identically distributed and the daily capacity for each bench is 5 hours.

Step 8.1. Complete hearing

- Task 1: Update the remaining daily capacity of the bench by subtracting the hearing time.
- Task 2: Randomly generate a hearing outcome and external delay for the associated hearing. If the outcome is "disposal" remove the case from the system. Otherwise, route the case to the back of an appropriate queue.
- Task 3: Remove the heard case from the cause list.
- Task 4: Go to Step 6 if the day is Monday or Friday. Otherwise, go to Step 7.
- Assumption: When a case returns to an appropriate queue, its position in the queue is based on the oldest-hearing first rule according to their arrival date to the stage (pre-admission and post-admission).
- Data source: We estimated the decision probability and the distribution of external delay from hearing order files. For the decision probability, we applied the decision tree algorithm (Section 3.3) to the order files. For the empirical distribution of external delay, we extracted key-words such as "list after 6 weeks" from the order files.

Step 8.2. Partial hearing

- Task 1: Include the part-heard hearing in the cause-list for the next working day. The hearing time for this part-heard case is its original hearing time minus the remaining daily capacity of the bench.
- Task 2: Update the remaining daily capacity of the bench to 0.
- Task 3: Go to Step 6 if the day is Monday or Friday. Otherwise, go to Step 7.

Step 9. Ending a day

- Task: Increase date by one and go to Step 1.

Step 10. Summer capacity activation

- Task: Check if the total backlogs are less than 32,000. If so, go to Step 9. Otherwise, generate a cause list of 60 post-admission cases (whose external delays are exhausted) and go to Step 11.
- Assumption: Only post-admission cases are processed during the summer holiday because they are the driver of the SCI's heavy congestion.

Step 11. Summer hearing

- Task 1: If the case is part-heard from the previous day, use the carried over hearing time as a hearing time for this part-heard case. Otherwise, sample a hearing time for the chosen case.
- Task 2: Compare the hearing time with the remaining daily capacity of the bench.

- Task 3: If the remaining capacity of the day exceeds the hearing time, go to Step 11.1. Otherwise, go to Step 11.2.

Step 11.1. Complete hearing

- Task 1: Update the remaining daily capacity of the bench by subtracting the hearing time.
- Task 2: Randomly generate a hearing outcome and external delay. If the outcome is "disposal" remove the case from the system. Otherwise, route the case to the back of an appropriate queue.
- Task 3: Remove the heard case from the cause list.
- Task 4: If the total backlogs are less than 32,000, return the remaining cases in the cause list to their original position in the queue and go to Step 9. Otherwise, go to Step 11

Step 11.2. Partial hearing

- Task 1: Include the part-heard hearing in the cause-list for the next working day. The hearing time for this part-heard case is its original hearing time minus the remaining daily capacity of the bench.
- Task 2: Update the remaining daily capacity of the bench to 0.
- Task 3: Go to Step 9.

Appendix B: Accuracy of our tree-based outcome classification model

Here, we report the confusion matrix from the accuracy testing discussed in Section 3.3.

Table 9 Confusion Matrix

	Actual →					
Prediction ↓	adjournment	disposal	regular	admitted	Total	Precision
adjournment	248	4	15	0	267	92.9%
disposal	2	104	7	3	116	89.7%
regular	5	2	93	0	100	93%
admitted	0	0	0	17	17	100%
Total	255	110	115	20	500	
Recall	97.3%	94.5%	80.9%	85%		

Appendix C: Additional Robustness Checks

C.1. Hearing time heterogeneity

In our main analysis, we assume that hearing times are independently sampled from a common distribution. We do this mainly because our hearing time dataset is not granular enough for us to separate hearing time distribution across different hearing types. Such an assumption might not be practical: for example, hearings that are eventually adjourned would have been much shorter than hearings with other outcomes; or, perhaps pre-admission hearing times are distinct from the hearing times for post-admission matters. This calls for robustness checks associated with our i.i.d. hearing time assumption.

C.1.1. Heterogeneity based on hearing outcome. Motivated by our observation that about 38% of hearings are adjourned, we separate hearing times by two groups. The first group keeps the bottom 38% of hearing times and the other goes to the second group. The average hearing time in the two groups are 0.87 minute and 10.46 minutes, respectively. In our simulation for robustness check, we randomly sample a hearing time from the first group if the underlying hearing is to be adjourned and do so from the second group otherwise.

The simulation confirms that our managerial insights derived from baseline remains valid with heterogeneous hearing times between adjourned hearings and non-adjourned hearings. In particular, with the 15/45 TWTh split, the system is in the quasi-stable regime and poor performance metrics (such as a high expected delay) can be alleviated by our intervention of rebalancing the TWTh split to 17/43. The results are summarized in Table 10.

Table 10 Key performance metrics with heterogeneous hearing times

	15/45 split (TWTh)	17/43 split (TWTh)
Expected delay	272.97 days	93.84 days
Excessive delay	17.6%	0.13%
Holiday hearings	1616 per year	0 per year
Forced idleness	19.06%	17.67%
Post-admission cases not reached (TWTh)	0.71 per day	1.01 per day
Pre-admission cases not reached (TWTh)	78.08 per day	84.05 per day

These results are very similar to those in Table 6 of the paper, hence, we conclude that our approach is robust. One might wonder why modeling a shorter adjournment hearing did not alleviate congestion. The reason is that the primary driver of congestion at the SCI is *forced idleness* of the judges. Because (by policy) the SCI does not schedule more than 60 cases in a day, if the hearings end early, then there is limited scope to hear additional cases. In other words, the benefit of shorter adjournment hearings is offset by increased forced idleness (compare Table 10 with Table 6).

C.1.2. Heterogeneity between pre- versus post-admission hearing. Although we have modeled a common hearing-time distribution, there could be potential differences between pre-admission versus post-admission hearings. However, it is unclear how different the two types of hearing times are. Our data does not allow us to cleanly distinguish between the two kinds of hearings. We further note that our understanding, based on direct observation of court proceedings and on interviews, is that pre-admission hearings can be quite similar to regular (post-admission) hearings. The reason is that as part of these hearings, substantial arguments about the merits of appealing a lower-court decision are presented.

Having said that, it is useful to demonstrate the robustness of our insights to the assumption that hearings are *iid*. Due to our inability to distinguish between pre- and post-admission hearing times in our data, we use a certain top percentile of the hearing time distribution to calibrate the post-admission hearing time distribution, and the remainder for calibrating pre-admission hearings.

In particular, we use the effective arrival rates to the different stages in the judicial queue (31,904, 41,447, and 24,167 for the fresh, after-notice, and post-admission matters, respectively to deduce that the proportion of post-admission hearings is 33%.[1] We then set aside the top 33% of the hearing times in our dataset to calibrate the post-admission hearing time distribution. This analysis yields an expected pre-admission hearing time of 1.46 minutes, and an expected post-admission hearing time of 17.54 minutes, a scaling factor of 12. The updated hearing time distributions are described in Table 11.

Table 11 **Summary statistics of pre-/post-admission hearing times**

Type	Min.	25%-tile	50%-tile	Mean	75%-tile	Max.
Pre-admission	0.82	0.87	0.88	1.46	1.77	3.48
Post-admission	3.5	5.15	8.55	17.54	18.46	271.08

Although a conservative robustness check is a good approach, such extreme heterogeneity is not present in the SCI data, which makes the calibration challenging. In particular, with this level of heterogeneity, we are unable to achieve a match with the first-order operational characteristics of the SCI, namely, the overall expected delay and number of holiday hearings. To do so, we need to relax the restriction to a 5-hour workday. With a 7-hour workday, we do get a baseline that matches the operational characteristics, as shown in the table below. The table also reports the outcome of our intervention based on rebalancing judicial capacity on regular hearing days (i.e., TWTh). For each simulation run, the choice of system parameters is summarized as a *configuration* represented as a 4-tuple whose first and second elements are the TWTh split across after-notice and regular matters, the third element is the threshold for triggering holiday capacity, and the final element is the workday duration in hours.

Table 12 **Robustness of our intervention with heterogeneous pre- and post-admission hearing time distributions.**

Configuration	Overall expected delay (days)	Excessive delay (%)	# holiday hearings/year
(45,15,32K,7)	268.95	17.5	1764.92
Baseline			
(44,16,32K,7)	260.04	17.7	396.49
(43,17,32K,7)	75.24	0.1	0

As can be seen above, the baseline split between after-notice (pre-admission) matters and post-admission matters on TWTh is (45,15). Altering this a bit to (43,17) alleviates the congestion in the post-admission queue (zero holiday hearings). This confirms the robustness of our insights.

C.2. Cross-validation of calibrated parameters in baseline simulation

To confirm robustness of the calibrated TWTh capacity split (15 post-admission cases and 45 after-notice cases) and holiday capacity activation threshold (32,000), we calibrate these from a smaller slice of raw data. The smaller subset covers arrival data from 2009 to 2013 and hearing time data from November 12, 2018 to February 12, 2019. (As this slice covers an earlier period of our raw data, one can consider our approach as

[1] See the "Analytical sufficient conditions for stability" paragraph in Section 4.3.1 for the computation of the effective arrival rates.

rolling window cross-validation.) The re-calibrated TWTh split is 15/45 and the holiday capacity activation threshold is 31,400, confirming that our initial calibration (15/45 split and activation threshold, 32,000) is robust. Efficacy of our intervention is also robust in the sense that after we change the TWTh split from 15/45 to 17/43, the SCI's operating regime shifts to underloaded from quasi-stable, and the key performance metrics are significantly improved, similar to our original analysis. We summarize the results in Table 13.

Table 13 Key performance metrics from re-calibrated parameters

	15/45 split (TWTh)	17/43 split (TWTh)
Expected delay	275.46 days	82.68 days
Excessive delay	17.57%	0.12%
Holiday hearings	1882 per year	0 per year
Forced idleness	16.81%	15.02%
Post-admission cases not reached (TWTh)	1.16 per day	1.75 per day
Pre-admission cases not reached (TWTh)	96.15 per day	108.35 per day

Appendix D: Ruling out congestion-driven admission control

While one expects congestion to not play a role in judicial decision making, this interaction is something that we look into in this section. As the congestion in the judicial queue varies (measured through backlogs), the court may modulate the admission probability into post-admission stage of the judicial queue.

We analyze the last hearing outcome (disposal/admission) in the pre-admission stage for a total of 191,937 cases. The outcome or dependent variable here, Y, is binary. It equals 1 if the case was admitted (to the second or post-admission stage) or 0 if the case was disposed. We consider the following set of covariates:

1. X_{backlog}: the number of outstanding cases in the SCI at the time when the outcome decision is made.
2. X_{year}: takes value 1 if hearing was conducted after 01/01/2015, otherwise 0. We track this variable because there was a big change in the acceptance rate into post-admission stage before and after 2015.
3. X_{holiday}: takes value 1 if hearing took place in a holiday month (a month with summer or winter vacation days), and 0 otherwise.
4. X_{duration}: counts the number of days the case has been in the system since arriving to the SCI. This controls for any potential time-in-system effects.
5. X_{future}: captures the average number of judges in the next 365 days. Due to planned retirements and new appointments, it is possible to anticipate changes in judicial capacity in the forthcoming year. This variable controls for this effect.

We use two estimation methods: logistic regression and decision trees. Table 14 displays the results of the logistic regression, and Figure 6 displays the results of the decision tree approach.

In Table 14, we observe that the coefficient $X_{backlog}$ is not significant and also the magnitude is extremely small. This reinforces our belief that the impact of congestion on judicial decision making in this setting is limited, if any. Next, look at the decision tree in Figure 6. In this figure, each node shows three numbers: (from top to the bottom) the predicted outcome where 1 represents "admitted", the predicted admission

Table 14 Results of regression analysis for admission control

Variable	Coefficient estimate	p-value
Intercept	-1.975	$< 2\times10^{-16}$
X_{backlog}	1.006×10^{-6}	0.579
X_{year}	1.143×10^{0}	$< 2\times10^{-16}$
X_{holiday}	3.739×10^{-1}	$< 2\times10^{-16}$
X_{duration}	1.244×10^{-3}	$< 2\times10^{-16}$
X_{future}	-5.635×10^{-1}	$< 2\times10^{-16}$

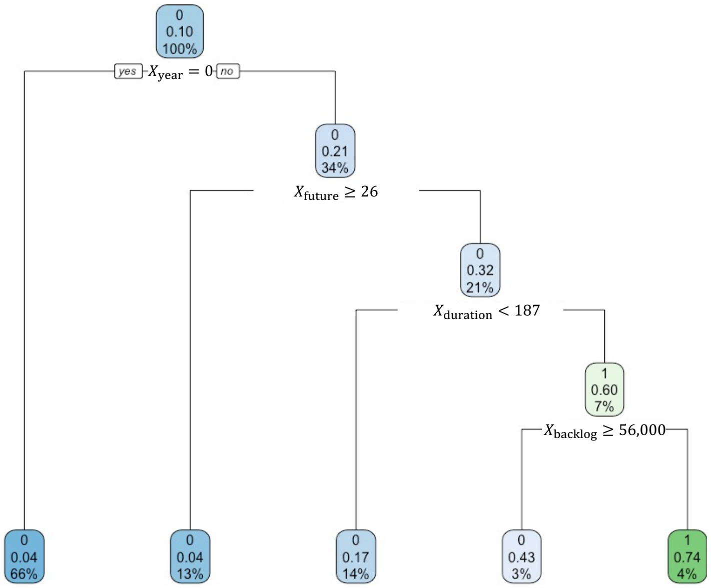

Figure 6 Decision tree for admission control.

rate in the note, and the percentage of observations in the node. The backlog does play some role here, albeit in only 7% of the cases. Further, we found that this model had an AUC of 80% as compared with 79.8% for the model without $X_{backlog}$. Thus, we believe the decision tree model also provides sufficient evidence that in our data, ignoring the impact of congestion on judicial decision making is not unreasonable.

Appendix E: Discussion about case heterogeneity: civil versus criminal matters

In this section, we present a quantitative discussion of considerations pertaining to case heterogeneity. Prior work indicates that more than 80% of the workload of the SCI pertains to dealing with appeals of two kinds: civil matters and criminal matters (Robinson 2013a). Therefore, we focus the discussion on civil versus criminal cases, and their relevant operational characteristics. Specifically, we discuss their respective hearing time distributions, and associated hearing outcome probabilities.

In the paper, we have modeled a single hearing-time distribution for all case types. Based on the display-board data, the expected hearing time for civil cases is 6.41 minutes and that for criminal cases is 6.23 minutes. The t-test cannot reject the null (p-value is 63%) that the two means are statistically different. Thus, our assumption that cases are homogeneous in hearing time is not restrictive.

We also report the estimated outcome probabilities by case type (i.e., civil versus criminal) and by year; see Tables 15 and 16 for pre-admission and post-admission decisions, respectively. We observe that except for certain years —2016 for pre-admission decisions, and 2011 and 2012 for post-admission outcomes— the difference in outcome probabilities is not substantial. Hence, as also explained in §5, after achieving stability via better scheduling, the additional impact of incorporating case heterogeneity on congestion will be small.

Table 15 Pre-admission outcome probabilities by case types per year

	Admission		Disposal		Adjournment		Regular	
Decision year	Civil	Criminal	Civil	Criminal	Civil	Criminal	Civil	Criminal
2010	0.39%	1.01%	46.80%	44.40%	27.70%	26.73%	25.07%	27.85%
2011	1.25%	1.84%	35.80%	41.21%	37.72%	32.21%	25.23%	25.74%
2012	1.59%	1.55%	34.88%	40.60%	38.64%	30.20%	24.89%	27.65%
2013	1.27%	1.51%	34.76%	37.05%	41.63%	35.20%	22.33%	26.23%
2014	1.27%	1.77%	37.47%	38.85%	39.71%	36.59%	21.55%	22.79%
2015	5.43%	4.72%	37.03%	38.57%	37.53%	34.92%	20.01%	21.80%
2016	12.08%	7.91%	31.01%	41.98%	39.53%	32.71%	17.38%	17.39%

Table 16 Post-admission outcome probabilities by case types per year

	Disposal		Adjournment		Regular	
Decision year	Civil	Criminal	Civil	Criminal	Civil	Criminal
2010	16.67%	12.50%	50.00%	50.00%	33.30%	37.50%
2011	11.43%	30.56%	70.48%	38.89%	18.10%	30.56%
2012	10.91%	21.85%	51.52%	50.42%	37.58%	27.73%
2013	12.50%	13.38%	55.31%	52.42%	32.19%	34.20%
2014	8.03%	3.59%	72.60%	81.59%	19.37%	14.81%
2015	17.73%	12.34%	57.06%	74.32%	25.21%	13.34%
2016	27.35%	21.87%	53.16%	58.57%	19.49%	19.56%

Appendix F: Confidence Intervals for Simulated Estimates

In this section, we report the size of 95% confidence interval for all the estimations given the main manuscript. For the three numbers, the expected delay in the pre-admission stage (74.65 days), the expected delay in the post-admission stage (1143.58 days), and the time until the first hearing in the second stage (1,124 days), reported in Section 4.2.2, the corresponding half-widths of 95% confidence intervals are 0.06 day, 0.13 day, and 0.09 day.

For Tables 5, 6, 7, and 8, the corresponding confidence intervals are reported, respectively, in Tables 17, 18, 19, and 20.

Table 17 Impact of Adding Judges

	14 benches		15 benches	
	Mean	Half-width 95% CI	Mean	Half-width 95% CI
Delay (in days)	275.46	0.42	252.75	0.38
Holiday hearings	1657 per year	495	84 per year	35

Table 18 Impact of Rebalancing Capacity between Stages

	15/45 split (TWTh)		17/43 split (TWTh)	
	Mean	Half-width 95% CI	Mean	Half-width 95% CI
Delay (in days)	275.46	0.42	95.6	0.09
Post-admission cases not reached per day (TWTh)	1.68	0.06	2.43	0.07
Pre-admission cases not reached per day (TWTh)	119.43	1.45	134.62	1.46

Table 19 Low Impact of Limiting Adjournments in Pre-Admission Stage

Reduction in Adjournment prob.	Expected Delay	Half-width 95% CI
5%	273.98 days	0.43 day
10%	274.87 days	0.45 day
15%	275.3 days	0.46 day
20%	272.37 days	0.46 day

Table 20 High Impact of Limiting Adjournments in Post-Admission Stage

Reduction in Adjournment prob.	Expected Delay	Half-width 95% CI
5%	92.44 days	0.09 day
10%	87.4 days	0.08 day
15%	84.74 days	0.08 day
20%	82.84 days	0.07 day