The Inefficiency of Worker Time Use*

Decio Coviello
HEC MONTRÈAL

Andrea Ichino
UNIVERSITY OF BOLOGNA

Nicola Persico
NEWYORK UNIVERSITY

December 9, 2011

Abstract

Much work is carried out in short, interrupted segments. This phenomenon, which we label *task juggling*, has been overlooked by economists. We study the work schedules of some judges in Italy documenting that they do juggle tasks and that juggling causally lowers their productivity substantially. To measure the size of this inefficiency, we show that although all these judges receive the same workload, those who are induced exogenously to juggle more trials at once instead of working sequentially on few of them at the same time, take longer to complete their portfolios of cases. Task juggling seems instead to have no adverse effect on the quality of the judges' decisions, as measured by the percent of decisions appealed. To identify these causal effects we exploit the lottery assigning cases to judges and the procedural prescription requiring the first hearing to be held within 60 days from filing. We view these findings as intriguing because task juggling implies that workers operate below their productivity frontier, at least as conventionally defined. We conclude discussing reasons why some judges are more inefficient than others.

JEL-Code: J0; K0; M5.
Keywords: Individual production function, work scheduling, duration of trials.

*We would like to thank seminar participants at the CEPR Public Policy meeting 2010, the NBER Summer Institute 2009, the WPEG 2009 conference, Bocconi University, European University Institute, LSE, MIT, Oxford, Royal Holloway, Timbergen Institute, Universities of Bologna, Bonn, Southampton UPF and Utrecht. We are also grateful to the Labor Court of Milan for making the data available to us and to Do Wan Kwak and Margherita Fort for sharing computer code with us and for useful suggestions concerning quantile estimation. Email: andrea.ichino@unibo.it; decio.coviello@gmail.com; nicola@nicolapersico.com

1 Introduction

The managerial literature on "time use" documents that workers frequently carry out a project in short incremental steps, each of which is interleaved with bits of work on other projects. For example, in a seminal study of software engineers Perlow (1999) reports that:

> "A large proportion of the time spent uninterrupted on individual activities was spent in very short blocks of time, sandwiched between interactive activities. Seventy-five percent of the blocks of time spent uninterrupted on individual activities were one hour or less in length, and, of those blocks of time, 60 percent were a half an hour or less in length."

Similarly, in their study of information consultants Gonzalez and Mark (2005, p. 151) report that:

> "The information workers that we studied engaged in an average of about 12 working spheres per day. [...] The continuous engagement with each working sphere before switching was very short, as the average working sphere segment lasted about 10.5 minutes."

The popular "self-help" literature has recognized that scheduling is a challenge for many workers. Books such as *The Myth of Multitasking: How "Doing It All" Gets Nothing Done* give workers suggestions to reduce multitasking on the job.[1]

The fact that much work is carried out in short, interrupted segments, a phenomenon which we label *task juggling*, has been overlooked by economists. It is strange because, at least in theory, task juggling directly reduces productivity. This is shown in the next example.

> **Example 1.** A worker is assigned two jobs, A and B, each requiring 2 days of undivided attention to complete. If the worker is exogenously induced to juggle both jobs, for example working on A on odd days and on B on even days, then the first task is finished after 3 days and the second after 4 days. The average duration is 3.5 days. If, instead, she is allowed to focus sequentially on each job in turn, then she completes A in 2 days and, later, B in 4 days from assignment. With this sequential work schedule the average duration is 3 days only and no

[1] The first two are: Resists making active [e.g., self-initiated] switches; and Minimize all passive [e.g., other-initiated] switches.(Cited from Crenshaw 2008, p. 89).

job is finished later. The difference between 3.5 and 3 is the mechanical effect of task juggling on average duration.

We study the work schedules of some labor judges in Italy. Our paper will document that, just as in the example, our judges juggle tasks; and show causally that juggling lowers their productivity substantially. These findings should give us pause in light of the assumption, common to most of academic economics, that workers operate on their productivity frontier. Task juggling, as we define it, carries no immediately apparent benefit for the worker (no effort is saved). If anything, it entails additional costs, beyond the effect described in Example 1, when switching attention between different tasks requires time and thus slows down job completion further. Nor is task juggling required of judges, as it may be of other workers in collaborative environments.[2] Finally, these workers are "experts." Of all workers, then, judges should be the least likely to be subject to this fallacy. And yet they fall prey to it. We take this as evidence that task scheduling is difficult to carry out in work environments, even when workers have controls over their schedule, even if they are experts, and even if the costs in terms of productivity are large.

Economists have paid little attention to this inefficiency because they typically have not considered the ability to schedule as a factor in the individual worker's production function. Only when scheduling is included as an "input" of this function it is possible to measure the size of the inefficiency. This is what we do, using an instrument which exogenously changes scheduling practices over time.

Three features of our environment are key to our estimation strategy. First, our workers (judges) operate essentially as single units: there is no team work involved in the production of their judicial decisions.[3] Secondly, we leverage the random assignment of cases to judges as a source of exogenous variation in the number and complexity of cases, the effects of which can be traced on the duration of cases. Finally, we are able to measure productivity, effort, ability and difficulty of tasks quite accurately.

We estimate the effect of task juggling on project duration by using an empirical specification derived from a theoretical model which generalizes Example 1. In our empirical setting, of course, judges cannot be expected to work on a single case at a time. But the number of cases they should efficiently be working on should be constant over time, and independent of

[2]Judges control their own schedule, and we show that they vary considerably in the amount of task juggling they engage in.

[3]One could argue that the lawyers are part of a team with the judge. However, the reality in our empirical setting is that judges have considerable authority over lawyers in limiting their possibility to slow down the trial. The constraint on completion time is judicial time, not lawyer time. Therefore the judge is to be considered as a single worker as regards completion time.

the rate at which cases are assigned. So if some cases are assigned earlier rather than later, the non-juggling judge should respond to this shock by keeping the "too early" cases on ice, and the effect on the duration of all other cases should be nil. But if the judge juggles, then he puts the newly assigned cases "in process" immediately and that has an externality on the duration of the other cases. In order to identify a "causal" effect we construct time varying instruments for effort and task juggling based on the sample realization of the lottery that allocates the amount and the typology of workload to each judge. This lottery is used in combination with the procedural rule prescribing that judges should hold the first hearing of a case no later than 60 days from filing. In this way, exogenous increases in the fraction of the assigned workload that reaches the "60 days" threshold, generate pressure for more task juggling.

Results strongly support the hypothesis that judges respond to an increase in caseload by juggling more tasks, and that this effect increases the durations of all cases. According to our estimates, an exogenously induced 8% increase of task juggling would need to be compensated by a 4% increase of effort in order to avoid an increse in the average duration of trials. Our results also suggest that the negative effect of increased task juggling is partly offset by an endogenous increase in effort on the part of judges (more hearings). Switching costs, on the other hand, appear to be less significant in our setting. We believe these estimates are the first empirical estimates of the impact of (inefficient) time allocation on productivity.

A comment on our measure of productivity. We focus on the duration of projects for two reasons. First, in many practical cases duration is what the worker's principals (clients) want to minimize.[4] Second, in our empirical application, reducing the duration of trials is a key statutory objective.[5] Duration of job completion is clearly not the only dimension of

[4]Many workers do not directly control the input in their productive process (such as when projects are assigned by the principal or by clients), but can control the speed at which their projects are completed. The latter tends to be especially true for workers who are not part of an assembly line. In these cases it is speed, for given quality, which is the relevant performance measure. For example, an IT consultant does not control the number of customers who need her services; when there is excess demand, increased productivity can only be achieved by reducing the duration of each job, from assignment to completion. In a different setting, whenever a contractor is hired, the principal (homeowner) cares about the speed of completion, for given quality.

[5]The Italian Constitution (art. 111) reads: "The law shall ensure the reasonable duration [of the trial]." And in (CSM 2010, p. 9), the Commission for the Setting of Standards in the Adjudication Process writes: "It is clear that, owing to the fundamental value attributed by the Constitution to the duration of trials, [...] a nationally-constructed index of duration must, sooner or later, become the standard measure of adjudication." At the European level there is a permanent Commission for the Efficiency of Justice (CEPEJ, see http://www.coe.int/t/dghl/cooperation/cepej) which is mainly focused on the duration of trials. At the global level, the "Doing Business" reports by the World Bank are concerned with the speed of dispute resolution.

output: quality matters as well. We will show that lower duration of trials is associated, if anything, with reductions in the probability that the judge's decision is appealed. Thus task juggling does not seem to generate any relevant trade off between quantity and quality for these workers, and we can focus on trials' duration only.

Although derived within the specific setting of Italian judges, the large estimated effect of time use practices on productivity that we estimate has more general implications and raises several important issues. First, it suggests that the managerial literature may be right in focusing on task juggling and time allocation. In fact, production functions which are estimated ignoring information on time use may be substantially misspecified, and sizable inefficiencies may be overlooked. Second, in our particular applications it raises the issue of the social cost of rules inducing more task juggling (the 60-days rule in the case of Italian judges). Third, the implications of our results apply also, more generally, to those situations in which more output is required, but labor or capital cannot be increased at least in the short run. A more sequential work schedule might offer a solution in these cases, because it increases output per unit of time at the cost of delaying the beginning date of some projects (but not their end date). This delay may not be optimal for other reasons in normal times, but may be the only feasible solution during workload peaks.

This paper fits broadly within the literature on the construction and estimation of production functions that can be traced back to the path-breaking article of Cobb and Douglas (1928).[6] Our goal is indeed to study and estimate the return to a factor of production but the focus is on individual (not firm) output. From this viewpoint, our results are more closely related to a recent literature initiated by Ichniowski et al.(1997), suggesting that, in different areas of human behaviour, individual modes of time use and activity scheduling are associated, in some cases causally, to performance for given effort.[7] Thanks to the accurate measurement of the steps of "production," and to the access to exogenous quasi-experimental variation, in this paper we are able to identify more tightly than in this literature the causal effect on productivity of a specific and well defined individual work practice, i.e., task juggling.

What we call task juggling is an inefficiency that is also related to the concept of "bottlenecks" in the literatures on project management and project planning (see Moder *et al.*,

[6] Jorgenson (1986) surveys extensively the origins of this literature.

[7] See, for example, Bertrand and Schoar (2003), Bloom et al. (2007,2009) and Bandiera et al. (2009) for CEO practices, Ameriks et al. (2003) and Lusardi and Mitchell (2008) for family financial planning and, closer to us, Aral et al. (2007) for multitasking activities and the productivity of single workers, and Garicano and Heaton (2010) for organization and productivity in the public sector. See also the recent surveys of Gibbons and Robert (2010) and Della Vigna (2009), the latter specifically on the issue of self-control in individual behaviour.

1983) and to the literature on network queuing, originating with Jackson (1963). We differ from the queuing literature in two ways. First, the queuing literature studies processes that are not explosive, meaning that a fraction of the time the queue is zero and the processor (the worker) is idle; this in not the case in our model, nor in our data. Second, the queuing literature is prescriptive: in our setting, it would prescribe to eliminate task juggling but it would not ask whether it prevails empirically.

Task juggling is also related to the sociological/management literature on time use.[8] This literature shows how frequent are working situations in which many projects are carried along at a parallel pace. Related to it is also the literature on the *disruption cost* of interruptions, surveyed by Mark *et al.* 2008. These literatures do not trace empirically the effect of task juggling on output, perhaps because individual output measures are hard to obtain in many work environments and also, presumably, because establishing a causal channel is challenging outside of an experimental setting. At a more popular level, there is a large time management culture which focuses on the dynamics of distraction and on "getting things done" (see e.g. Covey 1989, Allen 2001). The success of these popular books suggests that people do indeed find it difficult to scheudule tasks efficiently in the workplace.[9]

In our companion paper (Coviello et al. 2011) we address the question of why workers may want to engage in sub-optimal levels of task juggling; here, in Sections:why, we discuss what the answer to this question might be in the specific context of these judges. In the same companion paper we consider also the effect of different incentive schemes when the worker can arbitrage across projects of different complexity. Such arbitraging across tasks that cannot be individually incentivized has been called "multitasking" by Holmstrom and Milgrom (1991).

We present the data and the institutional framework in Section 2. Section 3 provides descriptive evidence on the correlation between the productivity of judges and their effort, their ability and their propensity to juggle tasks. In Section 4 we discuss the theoretical model that guides our econometric analysis, while Section 5 describes the econometric specification and the identification of the equations that we estimate. Results are presented in the same section, while Section 6 evaluates whether switching costs increase significanly the mechanical inefficiency of task juggling. Section 7, examines possible reasons why judges indulge in task juggling despite its inefficiency, and specifically what could cause the observed heterogeneity among them. Section 8 concludes.

[8] See Perlow (1999) and Gonzalez and Mark (2005) for examples and a review of the literature.

[9] For a review of the academic literature on this subject see Bellotti et al. (2004). For a specular take on prioritization of tasks see the discussion of the "firefighting" phenomenon in Bohn 2000 and Repenning 2001).

2 The data

We use data from one Italian court specialized in labor controversies for the industrial area of Milan. Our initial dataset contains all the 58280 cases filed between January 1, 2000 and December 31, 2005. For 92% of these cases we have information on their entire history, while the remaining cases are observed up to December 3, 2007. These trials have been assigned to 31 judges who have been in service for at least one quarter during the period of observation. For the judges who were already in service on January 1, 2000, we also have information on the cases that were assigned to them in the previous year and we can therefore compute a measure of their backlog at the beginning of the period under study. For the judges who took service during the period of observation (or less than one year before January 1, 2000) we analyze their productivity starting from the fifth of their quarters of service, in order to give them time to settle in. All the cases assigned to them during the first year of service (including those that were transferred to them from previous judges who left for another office or retired) are nevertheless counted to compute their backlog at the beginning of the second year of service in which we start to analyze their productivity. Thus all the judges that we analyze have at least one year of tenure, and for each we know the backlog of not-yet-disposed cases at the beginning of the period of observation.

We consider quarters as the relevant time unit and we focus on the subset of judges who received full workloads of new controversies within each quarter. We therefore eliminated the quarter observations concerning judges who did not receive a full workload because they retired, were transferred, were contemporaneously assigned to other duties or were in long term absence periods during which they were not receiving cases. At the end of this selection process, out of the original 31 judges we are left with the unbalanced panel of 21 judges described in Table 1. Six judges are observed for all the 24 quarters, while the others are observed for fewer quarters with a minimum of 8 quarters. The last column of the same table reports the number of cases assigned to each judge per quarter on average. The overall average is 128 cases per quarter-judge. The characteristics of the process that assigns cases to judges are crucial for the purpose of our study and require special attention.

In Italy, as in other countries, the law (Art. 25 of the Constitution) requires that judges receive a randomly assigned portfolio of new cases. This random assignment is designed to ensure the absence of any relationship between the identity of judges and the characteristics of the cases assigned to them. In the court that we consider the random assignment is implemented in the following way. Every morning the judges in service are ordered alphabetically starting from a randomly extracted letter of the alphabet. The cases filed during the day

are then assigned in alphabetic sequence to all judges in service. Note that this type of assignment scheme allows for small sample variability in the assignment of cases to judges, but this small sample variability is not systematic and fades away over the long run.

Table 2 shows, for example, that during the first quarter of 2000, the 18 judges in service received 129 cases on average with a standard deviation of 13 cases. The standard deviation is similar in all the other quarters. This because if, for example, in a given day the extracted letter is B and 5 cases are filed, only judges with a name starting from B to F will receive an assignment on that day (assuming one judge per letter of the alphabet). Therefore, within each quarter judges may receive slightly different workloads in terms of size.

For the same reason, also the characteristics of the assigned portfolios of cases may occasionaly differ across judges within a quarter. This is shown in the top part of Table 3 that reports, for each quarter, the p-value of Chi-square tests of indipendence between the identity of judges and three discrete characteristics of cases: type of controversy (14 types); zip code of the plaintif's lawyer (55 codes); the number of parties in trial (capped at 10). In the majority of quarters, independence cannot be rejected at standard significance levels, but in some quarters it is rejected at the 5% level. As shown in the second part of the table, this happens in 7 out of 24 quarters for the type of controversy, in 2 out 24 quarters for the lawyer's zip code and in 7 out of 24 quarters for the number of parties in trial. However, this occasional disomogeneity of the portfolios of cases assigned to judges fades away when the number of quarters over which judges are observed increases. This is shown in the last part of Table 3 that reports the p-values of similar Chi-square tests for all cases assigned in the period spanned by the largest balanced panel of judges identifiable in our sample. As verifiable in Table 1, this largest panel involves 14 judges observed continuously between year 2000 and year 2002. The p-values of these tests show clearly that independence cannot be rejected when we consider cases assigned over a sufficienly long number of quarters.

Therefore, we can conclude that, within a quarter, differences are due only to small sample variability and are not systematic. More specifically, they are independent of the identity of judges, who thus receive, in the long run, qualitatively and quantitatively similar portfolios of controversies. Note that, since our panel is unbalanced (see Table 1) we cannot test independence over all cases assigned to all judges in all quarters. Over the whole sample, independence is clearly rejected because judges with longer tenure receive larger numbers of cases and because different judges receive cases in different quarters and nothing guarantees the similarity of filed controversies over time. Nevertheless, the fact that independence cannot be rejected when we test over the largest balanced panel observable in our data, ensures that difference between all the judges observed in a quarter (even if they have

different tenure) are not systematically connected to the identity of judges, being due only to the alphabetic process of assignment described above.

As we will see in Section 5.1, for the purpose of identification of the causal effects of interest these are attractive and convenient features of our data that compensate for the unfortunate fact that we have no information of any kind concerning the judges under study, not even age and gender. Differently from other datasets, which typically have some demographic characteristics but do not contain measures of ability and effort, we instead observe the entire history of all the cases assigned to each judge. With this information we can construct, as we will see in the next section, very precise time-varying measures of productivity, work scheduling, ability, and effort for each judge.

3 Descriptive evidence

In this Section, we compare judges on the basis of average indicators of performance per quarter, computed over all the quarters in which each judge is observed.

3.1 Total duration and active cases

The height of circles (marked by the judge id number) on the vertical axis of the top left panel of Figure 1 measures the total duration of cases assigned to each judge. Total duration is defined as the number of days from filing until the date in which a sentence is deposited by the judge, or the case is settled, or censoring occurs in the few cases for which we do not see the end of the trial.[10] On the horizontal axis judges are ordered from the slowest one to the left (Judge 30) to the fastest one to the right (Judge 3). The height of the squares in the same panel indicates the workload of new cases assigned to each judge on average per quarter. This graphic representation makes transparent the heterogeneity of performance, in terms of duration of trials, observed for these judges despite the fact that they receive a workload which is fairly similar in quantity (because we selected only judges who receive a full workload) and quality (because of random assignment). For example, at the opposite extremes, Judges 30 and 3 receive respectively 120 and 105 cases per quarter, but the first one needs 398 days to close them while the second one need only 178 days, i.e., less than half.

The bottom left panel in the same figure plots the number cases on which each judge is contemporaneously working on average in a quarter. We call these "active" cases and they will be the focus of our analysis because they measure the extent to which judges practice

[10] See Section 2.

task juggling. Formally, a case is defined as active at a given date if its first hearing has already taken place but the case has not been completed yet. Of course we do not know the exact moment in which a judge starts working on cases previously assigned to her, but it seems reasonable to consider the first hearing as a good approximation of this moment. Also in this panel (as in all the others of this figure) judges are ordered from the slowest one on the left to the fastest one on the right. The vertical comparison between the left panels of the figure highlights the striking correlation across judges (0.93) between the average number of active cases and the average duration of trials. Comparing again extreme cases, the slowest Judge 30 keeps on average 275 files contemporaneously open on his desk while Judge 3 works on only 116 cases at the same time. In general, those who "keep more pots on the fire need more time to complete meals". It is important to keep in mind that these differences emerge among judges of the same office, who work in exactly the same conditions, with the same secretarial assistance and with a similar workload in terms of quantity and quality.

3.2 Throughput and backlog

For the reasons explained in the introduction, we prefer to use duration as opposed to throughput as a measure of productivity. The total throughput of these judges can only be equal to the input they receive, in terms of cases exogenously assigned to them. In principle, two judges may be deciding the same number of cases in a given quarter, but for one of them these cases may have been assigned just recently while for the other they may be very old cases. What matters, really, is how long it takes to process the input. Nevertheless, Figure 1 shows that the two measures are correlated. More precisely, it shows that if keeping too many files opened at the same time slows down the activity of a judge, also the number of cases he will be able to close per quarter will be negatively affected on average (but not necessarily within each specific quarter as we argued in the introduction). The top central panel of Figure 1 confirms this intuition by plotting the throughput of judges ordered, as usual, from left to right according to speed of case completion. The slowest Judge 30 has almost the worst throughput (106 cases per quarter, just 8 more than the worst performer, Judge 29). The most productive in terms of throughput is Judge 11 (131 cases per quarter) who is the second best performer in terms of duration. The correlation between the number of active cases and the number of closed cases across judges per quarter is -0.36 and suggests that judges who work on few cases at the same time, opening new ones only when older ones are closed, can not only dispose of assigned cases in less time from assignment but also increase their throughput per quarter.

Consistently with this hypothesis, it is not surprising to infer, from the bottom central

panel of Figure 1, that the fastest judges with fewer active cases have on average a lower backlog at the beginning of each quarter. This backlog ranges from the 545 cases of Judge 18, who keeps 258 cases open at the same time and is one of the worse performers in terms of duration and throughput, to the 230 cases of the already mentioned top performer Judge 3, who has on average only 116 files on his desk at the same time. Even if all these judges receive the same number of cases per quarter their backlog is highly correlated with the number of active cases (0.94).

3.3 Complication of cases, ability and effort of judges

Although suggestive, our hypothesis concerning the role of task juggling on the productivity of judges must be confronted with other more obvious potentially relevant determinants of this performance. In particular, ability and effort.

Consider the average number of hearings that a judge needs to close a case. Without random assignment this statistic would depend on both the difficulty of the cases assigned to a judge and on her ability to handle them quickly. But given random assignment, the complication of controversies that judges face should be fairly similar, up to small random differences determined by the realization of the assignment procedure described in Section 2. Therefore, differences across judges in the average number of hearings to close a case should mostly capture the unobservable skills that determine how a judge can control the trial and the behaviour of parties, lawyers and witnesses, in order to reach quickly a decision.

This statistic is plotted in the top right panel of Figure 1, where judges are again ordered, on the horizontal axis, from the slowest one on the left to the fastest one on the right. In contrast with the previously examined panels of this figure, here we do not see a clear pattern jumping out of the data. Some slow judges on the left (like 30 and 18) require less than 3 hearings to close a case on average, while many faster judges need more (including in particular the top performers 3 and 14). The correlation between duration and number of hearings per case is positive (0.18) but relatively low. Inasmuch as being able to decide a case with fewer hearings is a form of ability of a judge, this descriptive evidence does not suggest that such characteristics has a strong effect on performance as measured by total duration of cases.

A measure of effort is instead offered in our data by the number of hearings per unit of time. The idea is that, by exerting more effort, a judge can schedule more hearings per quarter and in this way can ceteris paribus improve her performance in terms of throughput and total duration of completed cases. This statistic is plotted in the bottom right panel of Figure 1 and also in this case we cannot infer an evident pattern connecting this measure of

effort to performance in terms of duration (the correlation is -0.06).

To summarize, the descriptive evidence presented in this section suggests that task juggling, as opposed to sequential working, may reduce considerably the performance of judges in terms of throughput and total duration of the cases assigned to them. Indicators of experience, ability and effort are as well likely to be relevant determinants of performance, but in a possibly less significant way. However, to properly assess the relative importance of these factors a theoretical framework and a multivariate statistical analysis are needed, to which we turn in the next Sections 4 and 5.

Before doing so, it seems important to say a word on the possibility of a "quantity versus quality" trade off in the performance of judges. Could it be that the judges with the highest throughput and the lowest total duration are worse judges in terms of quality of decisions? The evidence presented in Figure 2 suggests that the answer is no, as long as the percent of appealed cases can be considered as a good measure of the quality of the judges' decisions. There is no evidence that the cases assigned to slow judges on the left have a lower probability of appeal than the cases assigned to fast judges on the right. If anything the opposite seems to hold, given that the correlation between total duration and the percent of appealed cases is positive (0.41). For this reason we focus just on the effect of task juggling on duration in the rest of this paper.[11]

4 A theoretical framework to estimate the inefficiency caused by task juggling

In this section we seek a theoretical expression for the production function of judges that we will then use as a basis for the estimation of the inefficiency caused by task juggling. The measure of output we focus on is the duration of cases.[12] We will derive an expression for duration of a case based on the effort put in by the judge, the complexity of the case, and the way that the judge organizes its work schedule. The latter input is the novelty of our

[11] We have performed an empirical analysis similar to the one described in Sections 5 and 6, using the percent of appealed decisions as the dependent variable, instead of the duration of cases. This analysis shows that task juggling does not seem to have any adverse effect on the quality of the judges' decisions, as measured by the percent of decisions appealed. Results are omitted to save on space but are available from the authors.

[12] An alternative measure of productivity could be the sum of disposed cases in a given time interval. For the reasons detailed in the introduction this is not a straightforward measure of productivity across time in this and many other contexts. Note, in particular, that judges do not control their assigned workload but only the speed at which they complete it, which translates into a cumulated sum of disposed cases at any given quarter from assignment. Therefore, the average duration of job completion and the *cumulated sum* of disposed cases from assignment are two equivalent output measure for these judges.

analysis, and is measured by the number of cases the judge is working on at the same time.

In this section we study this production function in a steady state, where the number of cases the judge is working on at the same time is constant. Then, in Section 4.3 we study the effect of a perturbation around this steady state: i.e. an exogenously induced change in the timing at which a judge opens the cases assigned to her. The instrument that we will use in our empirical analysis mimic this exogenous perturbation.

To simplify matters, we start by assuming that there are no switching costs; that is, no time wasted when a judge switches from one trial to another. In other words, there is no loss of output related to the need of refocusing attention. We will introduce this complication later in Section 6 in order to assess its relevance.

4.1 The basic setup

The effect of effort and complexity can be appreciated even in the most stark model in which only one case is assigned to the judge. This situation is particularly simple because there is no question of how effort is distributed among different cases. The only factors that determine duration, then, are the number of hearings, or steps, that it takes to adjudicate the case (which we denote by S) and the number of hearings the judge makes per quarter (which we denote by e_q). Under the assumption that the judge exerts the same effort in every quarter we have $e_q = e$ and thus the duration of the (single) case has a very simple expression:[13]

$$D = \frac{S}{e}. \tag{1}$$

A similar expression can be derived when e_q is not constant across quarters.

When the judge receives more than one case, a third factor beyond e and S affects the duration of cases, namely, how the judge allocates his time across different cases. To describe the possible ways in which this is done we need to develop some language. This language, and the associated mathematical results, are formally developed in the Appendix Section 9. Here we describe intuitively several possible modes of work, that is, several algorithms that the judge may use to allocate his time across different cases.

First is the sequential work schedule, in which cases are worked on sequentially: first all the steps relating to the first assigned case are accomplished, then all the steps relating to the second assigned case, etc. The polar opposite of a sequential work schedule is the full rotation one, in which, within each step, all cases that the judge has received are worked on simultaneously until each is done. A partial rotation is a generalization of the previous

[13] Actually, to be precise the duration would be the smallest integer that exceeds S/e, but from now on we will ignore such integer problems.

process: it works just as a full rotation does, except that instead of rotating on all cases the judges has received, the judge keeps some cases unopened and only gradually inserts them into the rotation. Once a case is inserted in the rotation then it receives the same amount of attention as all other open cases. We define the opening of a case as the action of inserting that case in the rotation of all the other cases that are already opened.

The full rotation and sequential work schedules are polar extremes. In the full rotation schedule cases are started as early as possible and so, at any given point in time, there is a large mass of cases being simultaneously worked on. In contrast, a sequential work schedule causes the start of a new case to be postponed as late as possible, and so in a sequential work schedule the minimum possible number of cases is simultaneously being worked on at any point in time. The partial rotation is a general family of work schedules which subsumes as special and extreme cases the full rotation and the sequential schedule. This family is parameterized by the distance from assignment at which cases are opened. That is to say, a partial rotation can take different forms depending on how early after assignment the cases are opened. If, for example, all cases are opened as early as possible (i.e. immediately after assignment) then a partial rotation becomes identical to the full rotation. If, instead, new cases are opened at the slowest possible pace, then there is only one case open at any given time and so the partial rotation coincides with the sequential work schedule. This is why a partial rotation open is a convenient family of work schedules to work with.

Having introduced the notion of a partial rotation, we now want to use it to describe how time allocation practices affect productivity. In other words we want to generalize equation (1) by introducing a parameter which captures how close a partial rotation is to its polar extremes. The challenge here is that the production process we study evolves over time. To meet this challenge it is easiest to start by focusing on a system that evolves in a stable way as far as the number of cases included in the rotation is concerned. To this end we now introduce the simplest possible evolution of the system over time.

Definition 1. *A judge operates according to a* ***stable rotation*** *if:*

(a) in each quarter the judge keeps A_0 open cases;

(b) the number α of cases assigned, their complexity S, the effort e, and the number of new opened cases ν, are all constant in each quarter;

(c) the work schedule is a partial rotation;

(d) the number of cases completed ω is constant across quarters, and is the same as the number of new opened cases ν.

A stable rotation describes the production process of a judge who works according to a

partial rotation on a number of cases which remains stable over time. Figure 3 describes a snapshot of a judge's caseload in a stable rotation. Each folder represents a case and the horizontal axis is the number of hearings (steps) of that case that have already been completed. In this example, each case requires $S = 5$ hearings to complete. At the time of the snapshot, this judge has 5 opened cases that have had one hearing, 5 opened cases that have had two, and so on. Cases which are closer to completion are colored in a lighter shade. To the left of the vertical axis are cases which have not yet been started. The white folders represent cases that are done, i.e., have received 5 hearings.

Starting from this snapshot, if we let time run forward we will see that the judge holds one hearing for every opened case; this is because the judge follows a partial rotation. Graphically, this effort moves all folders one step to the right. In addition, the judge opens the five cases to the left of the vertical axis. Let us imagine that this is all the effort the judge has time for in a quarter (this implies $e = 25$). In this case $A_0 = 20$, and the input rate is exactly equal to the throughput rate, as it must be in a stable rotation. The throughput in a quarter is exactly 5 cases, which is equal to e/S. This equality is no coincidence: in Appendix 9.1 we prove that the input rate and the output rate must be exactly equal to e/S for there to be a stable rotation.

This finding shows that ν, our measure of congestion and time use, is not a free parameter in a stable rotation. Therefore we will have to go beyond a stable rotation if we want to examine the effect of a change in time use. This will be done in Section 4.3.

Note that in a stable rotation the duration of cases D_q need not be constant over time. Indeed, in a stable rotation the backlog of cases will grow if the arrival rate of cases exceeds the rate at which they are opened. In Appendix 9.1 we fully analyze how a stable rotation operates and obtain the following expression for the duration of cases.

$$D_q = \frac{S}{e}\left(A_0 + \alpha q\right) - q. \tag{2}$$

This expression solves for the duration D_q of cases assigned in quarter q in terms of known quantities: the exogenous assignment rate α, the measure of effort e/S, and the initial condition A_0, which is a parameter that can be specified arbitrarily. If a judge starts out with A_0 active cases in $q = 0$, and new cases are opened at the rate of e/S in quarters $q = 1, 2, ...$, then cases will be solved at a rate of e/S per quarter and at all times there will be A_0 active cases. While the output rate of cases does not depend on A_0, the duration of each individual case does according to expression (2).

Using this expression we can illustrate some of the determinants of duration, albeit at a stable rotation. The duration of a case is increasing in α, the rate at which cases are

assigned to the judge. It is decreasing in e/S, which means that judges who work hard (high e) or who have easy cases and/or are more able (low S) will have a lower duration of cases in steady state. Having a large number of active cases A_0 increases duration. Finally, the duration of cases increases with the judge's tenure ($\frac{\partial D_q}{\partial q} > 0$) if and only if $\alpha > e/S$, that is, if the arrival rate exceeds the judge's effort scaled by the perceived complexity of cases. We record these findings in a proposition.

Proposition 1. *If judges operate according to a stable rotation, the duration of a case assigned at q is increasing in α, in S/e, in A_0 and, if $\alpha > e/S$, also in q.*

Equation (2) and Proposition (1) provide a theory-based starting point for implementing an econometric analysis of the contributing factors to durations. While in standard theories of the individual production function, that ignore the scheduling of tasks, the duration of trials would depend only on the size of the workload, the difficulty of cases, the effort and the ability of a judge, our framework suggests, instead, that how time is allocated across cases for given effort and ability must be included in the specification. The natural way to test the proposition would be to estimate a linear approximation of equation 2:

$$D_{i,q} = \gamma_0 + \gamma_1 \alpha_{i,q} + \gamma_2 \left(\frac{e}{S}\right)_{i,q} + \gamma_4 q + \gamma_5 A_{i,0} + u_{i,q} \tag{3}$$

where $D_{i,q}$ is the duration of cases assigned to judge i in quarter q, $\alpha_{i,q}$ is the number of these cases (the workload), $\left(\frac{e}{S}\right)_{i,q}$ is effort standardized by the complexity of cases as perceived by the judge (which is also, potentially, a measure of ability), q is a time trend, $A_{i,0}$ is the initial judge-specific condition that defines the stable number of cases on which the judge rotates tasks. The presence of the error term $u_{i,q}$ is justified because in the data the workload, effort and complexity are not constant over time, while, if they were constant, equation 2 would be an exact relationship.

However, this specification is unsatisfactory because the congestion and time misallocation induced by task juggling are constant in a stable rotation. Since we want to estimate the effects of changes in congestion, this is a problem. From an empirical viewpoint, as well, the judges are observed to depart, albeit slightly, from the model of a stable rotation. The next Section explores the extent to which a stable rotation model captures the behavior of our judges.

4.2 Are judges scheduling tasks according to a stable rotation?

To establish whether judges effectively work according to a stable rotation we have estimated a regression of the number of open cases ν on the number of closed cases ω, obtaining the

following results:[14]

$$\nu = \underset{(5.55)}{5.99} + \underset{(0.04)}{1.01}\,\omega \tag{4}$$

where standard errors are reported in parentheses under the coefficients. According to these estimates these judges work on a schedule that is very close to a stable rotation but does not coincide exactly with it. The slope is approximately equal to 1 indicating that judges open one new case for each case that they close. But the positive intercept (even if statistically not significant) suggests that on average they also open approximately 6 new cases in every quarter on top of those that they close. As a result the number of active cases on their desk steadily increases over quarters albeit at a relatively low pace.

This pattern can be appreciated graphically in Figure 4. The top left panel plots the number of cases opened and closed per quarter by the seven best judges in terms of average duration. The two lines are very close one to the other, which is what should happen if these judges work according to a stable rotation, but the numbers of opened and closed cases, albeit similar, are clearly not constant overtime. The top right panel repeat the exercise for the seven worst judges. For these judges it happens more frequently that the number of new opened cases is larger than the number of closed cases. It is therefore not surprising to find, in the bottom left panel, that the seven worst judges have more active cases in each quarter. This panel also shows that for both type of judges (and in particular for the worst) the number of active cases increases over time with jumps that obviously correspond closely to the quarters in which more cases are opened than closed. Finally the last panel shows that the duration of all assigned cases differs across the two groups of judges and evolves over time within each group, in line with the number of active cases, as predicted by our model.

This evidence suggests that some judges are closer than others to a stable rotation schedule. But deviations from a stable rotation exist (in both directions) and have important effects on the number of active cases and on the duration of assigned cases. Thus a stable rotation is limited in its ability to account for what we see in the data and more generally to explain what is the effect of an *increase* in congestion. Indeed, in a stable rotation the amount of congestion is constant because, by definition, cases are opened at the same rate at which they are completed. We will therefore generalize our framework in the next Section 4.3, to the more interesting and realistic case in which congestion can change.

[14]The regression has been estimated on 381 quarter-judge observations and include fixed effects for the 21 judges.

4.3 The effect of an induced changed in task juggling

In this section we derive and sign the effects on durations of changing the timing at which cases are assigned to the judge. If a batch of cases is assigned earlier rather than later, we find the duration of all cases increases. According to our theory, this is because the judge juggles tasks, and so puts the batch of newly assigned cases "in process" immediately. Note the contrast with the effect we would expect if the judge was not juggling; in that case the order with which he gets to cases would be independent of the time at which cases are assigned, and so the duration of cases would be, too.

Therefore, we now consider the effect of an exogenous shock that induces judges to increase by Δ the number of cases ν_q newly opened in quarter q, relative to a stable rotation where cases are opened at the constant rate ν. Figure 5 describes this event. Remember that, by definition, the opening of a new case consists in inserting it into the rotation that involves all the other already opened cases. In the following proposition we assume, for convenience and without loss of generality, that cases are opened immediately upon being assigned.[15]

Proposition 2. *Suppose that there are no switching costs and that the judge operates according to a partial rotation. Suppose that a judge has so far opened ν cases per period and has exerted constant effort in each period. Suppose now that, for a given q only, the judge makes $\nu_{q-1} = \nu + \Delta$ and changes nothing else. Then the duration of cases assigned at q increases.*

Proof: See Appendix 9.2 ■

This proposition extends the introductory example of Section 1 to the case in which the judge works on an infinite stream of cases. Like in that example we find that if more cases are opened sooner, the average duration of all cases increases. To test this proposition, that applies to judges who deviate occasionally and for exogenous reasons from a stable rotation, the econometric specification (3) must be corrected to include a variable $P_{i,q}$ measuring how task juggling changes with respect to the initial condition. We use two alternative but related variables to do so. First, a flow variable: the number of new opened cases by judge i in quarter q, denoted by $\nu_{i,q}$. Second, a stock variable: the number of active cases on the table of judge i at the end quarter q, denoted by $A_{i,q}$. Thus, the specification that we want

[15] The convenience in the assumption is that we do not have to keep track of the time a case spends assigned but unopened. Nothing would change in the Proposition if we assumed that all assigned cases wait a fixed amount of time, say 60 days, before being opened.

to estimate is:

$$D_{i,q} = \beta_0 + \beta_1 \alpha_{i,q} + \beta_2 \left(\frac{e}{S}\right)_{i,q} + \beta_3 P_{i,q} + \beta_4 q + \delta_i + \epsilon_{i,q} \tag{5}$$

where δ_i is a judge specific fixed effect that absorbs the initial condition $A_{i,0}$, even if it is not observed for some judges.

What signs does the theory predict for the coefficients in this relationship? The signs of β_1 and β_2 are almost predicted by Proposition (1), but not exactly since Proposition (1) deals with the case of a permanent change in $\alpha_{i,q}$ and $\left(\frac{e}{S}\right)_{i,q}$, whereas β_1 and β_2 measure the effect of a temporary increase in their respective variables. So, for example, β_1 measures the effect on duration of going from $\alpha, \alpha, \alpha, \alpha, ...$ to $\alpha, \alpha, 2\alpha, \alpha, ...$To establish the signs of β_1, observe that an increase in $\alpha_{i,q}$ means that more cases are exogenously assigned to judge i in quarter q. Therefore, when the time comes for the judge to work on these cases, it will necessarily take longer to complete them whatever the scheduling of tasks chosen by the judge. Most theories of the duration of trials, would predict, like ours , that $\beta_1 > 0$.[16,17]

Perhaps less controversial is the prediction that $\beta_2 < 0$, because an increase in standardized effort $\left(\frac{e}{S}\right)_{i,q}$ means that the judge holds more hearings in quarter q (for whatever cases are open on her desk), or reduces the number of hearings $S_{i,q}$ needed to close the cases assigned to her. $S_{i,q}$ increases $D_{i,q}$ mechanically, because it means that cases assigned in q are more complex (or are considered as such by the judge), and so they take more tasks to adjudicate. Note also, as discussed in Section 2, that within each quarter, by random assignment, all judges receive portfolio of cases that should differ just because of random sampling. Therefore, if $S_{i,q} > S_{j,q}$ it must be either because judge i has randomly received a slightly more complex portfolio, or because the portfolio is effectively identical but judge j is "more able" in the sense that she can close the same portfolio of cases with fewer hearings on average than judge i. Moreover, for the same judge across quarters, it could happen that $S_{i,q} > S_{i,p}$, with $q < p$, and this may happen either because the ability of judge i increases over time or because the assigned cases becomes less difficult on average over time.

[16]But in the presence of learning by doing, economies of scale or positive externalities between cases, one could imagine that a larger workload might reduce the average duration of assigned cases. We will deal with such considerations later.

[17]Note that if the workload $\alpha_{i,q}$ were exactly equal for all judges within each quarter, the inclusion of judges' fixed effects and quarter fixed effects, on which we come back below, should prevent the identification of β_1 because of multicollinearity. But as explained in Section 2, cases are assigned to judges in order of arrival on a daily basis by alphabetical order, starting with the judge whose letter is extracted in the morning. So, if there are 10 judges in service and 15 filed cases, five judges will receive 2 cases and the other five only 1 and in the following day the assignment procedure restarts from scratch with the extraction of a new letter. The assignments may therefore differ slightly across judges but in a way that is uncorrelated with any non-ignorable characteristics of judges. See Section 2 and specifically Table 3. Thus, even controlling for quarters and judges fixed effects, the data display judge specific variability over time of the workload $\alpha_{i,q}$.

The main focus of our analysis is on the parameter β_3 which measures the inefficiency of task juggling, i.e. its effect on the average duration of all trials assigned in a quarter. Proposition 2 states without ambiguity that this coefficient should be estimated to be positive independently of whether task juggling is measured by $P_{i,q} = \nu_{i,q}$ or $P_{i,q} = A_{i,q}$.

Finally, Proposition (1) gives the condition for the coefficient on the time trend β_4 to be positive. We specify this trend in the most flexible way as a set of dummies for each quarter, so that we can control also for seasonality, and we expect the trend implicitly defined by the quarter dummies to be positive.

5 Estimates of the effect of task juggling on the average duration of trials

While $\alpha_{i,q}$ is randomly assigned (see Section 2), if work scheduling has a role in the determination of the duration of trials the error term $\epsilon_{i,q}$ in equation 5 is correlated not only with standardized effort $\left(\frac{e}{S}\right)_{i,q}$ but also with the amount of task juggling $P_{i,q}$, however measured. This because the error term includes lagged and forward values of standardized effort as well as the unobservable component that captures judge specific preferences concerning task juggling. There is, in principle, no reason to expect that this parameter should be time invariant.

Therefore to estimate consistently the causal effects of standardized effort and task juggling on trials duration with equation (5), we need some exogenous source of variation of these two variables.

5.1 Identification

As far as standardized effort is concerned, this exogenous source of variation is offered by the alphabetical system, discussed in Section 2, that determines the assignment of cases to judges on a daily basis. As a result of this system, within a specific quarter judge i may receive a slightly larger fraction of urgent or complicated cases than judge j, simply because of the randomly chosen letter of the alphabet from which the assignment of cases to judges was started in the days of that specific quarter. We therefore use as instrument for standardized effort the fraction of "urgent" cases and the fraction of "difficult" cases that judges receive in each quarter.[18]

[18]The classification of a case as "difficult" was implemented using an independent survey of judges whom were asked to classify the typology of possible cases according to their complication. "Urgent" cases are instead those cases that by law have to be completed in one hearing to be held almost immediately after filing, for example because some crucial worker's right is under prejudice and immediate protection is needed.

Note that these instruments, which capture the complexity or urgency of assigned controversies conditionally on the size of the workload, affect the duration of cases mainly through the effort e or the ability/ perceived difficulty of cases S. For example, if judge i receives randomly more difficult cases than judge j in a given quarter, this event can affect duration only if the judge changes the number of hearings per quarter ($e_{i,q}$) or if he changes the number of hearings needed to adjudicate the cases assigned in the quarter ($S_{i,q}$).

An instrument for the amount of task juggling $P_{i,q}$, whether measured with the number of new opened cases $\nu_{i,q}$ or with the number of active cases $A_{i,q}$, can instead be constructed exploiting a procedural prescription that constraints the freedom of judges to decide when to hold the first hearing of non urgent cases. Judges are in fact invited to hold the first hearing of these cases within 60 days from filing. There is no penalty for a delay but if long delays become systematic the judge may be put under disciplinary investigation by the *Consiglio Superiore della Magistratura*, i.e., the independent body that governs judges. As a result of this prescription, if the number of non-urgent cases assigned to judge i increases in the current quarter, the number of cases reaching the "60 days" threshold in the next quarter will be higher, putting some pressure on judge i to open more new cases. Descriptive evidence concerning the pressure generated by the "60 days" rule on judges is offered in Figure 6, which plots the distribution of inactive duration, i.e. the number of days between assignment and the first hearing of non-urgent cases. The figure suggests that judges rarely touch cases before they are "late," i.e. before 60 days from assignment. This is presumably because they are busy opening other, more ancient cases. After a case is already late, then a judge feels the pressure to try to open it soon in order to minimize the days of violation of the "60 days" rule. This behavior would be consistent with the notion that it is the most egregious violations of the "60 days" rule that might get the judge in trouble. In other words, the "60 days" rule works essentially like a bell that rings reminding judges that they should start acting on cases. Penalties for trespassing hit only if judges wait "too long" to react after the bell.

We therefore construct an instrument $Z_{i,q}$ for the amount of task juggling $P_{i,q}$, defined as the ratio of the number of cases assigned to judge i in the previous quarter divided by the total number of cases assigned in the previous and current quarters:

$$Z_{i,q} = \frac{\alpha_{i,q-1}}{\alpha_{i,q-1} + \alpha_{i,q}} \tag{6}$$

This instrument captures the idea that judges who feel the pressure of the "60 days" rule will open more new cases and this is expected to increase the duration of all their

These cases typically anticipate the related underlying trial that follows later as a separate case.

assigned cases. However, as any assignment-to-treatment mechanism, also this one suffers the possibility of non-compliance. Not all judges feel the pressure of the "60 days" rule, but some do feel it, as suggested by Figure 6 and by the first stage statistics discussed in the next section 5.2, and tend to open more or less new cases depending on which fraction of trials, within the recently assigned load, gets near or has just passed the "60 days" threshold. Note that, as explained in Section 2, the instrument is randomly assigned because it depends only on the assignment of new cases to judges in the current and previous quarters, which results from the propensity to litigate of workers and firms in Milan and from the alphabetical daily assignment system. Moreover, as we show in the next section, the small sample variability generated by the alphabetical daily assignment system ensures that the instrument is sufficiently strong. It also displays judge specific variability over time and is therefore compatible with the inclusion of judges and quarters fixed effects.[19]

Anecdotal evidence on the relevance of this instrument is offered by the fact when the results of this research were made public in Italy, some judges who had been previously put under investigation because too many of their first hearings took place far beyond the "60 days" threshold, informed us by email that they defended themselves showing that, by working sequentially, they had lower average durations than their colleagues. And were indeed acquitted on the basis of this evidence, which is completely in line with the prediction of our theory.

5.2 Estimates

The goal in this section is to verify whether the estimated coefficients of equation 5 have the sign predicted by the theory. If they do, then we cannot reject the hypothesis that these judges work according to a partial rotation, i.e., they juggle tasks. The magnitude of the coefficient associated with an increase in juggling will measure the inefficiency induced by task juggling.

Table 4 gives the descriptive statistics for the variables used in the econometric analysis, while results of the estimation of equation 5 are presented in Table 5. In the first column the amount of task juggling $P_{i,q}$ is measured with the number of new opened cases per quarter $\nu_{i,q}$, i.e., the number of assigned cases for which the judge holds the first hearing in the

[19]Note also that $\alpha_{i,q-1}$ would be an alternative simpler instrument and it would be randomly assigned as well. However it would violate the exclusion restriction because a higher workload in the previous quarter delays completion of cases assigned in the current quarter not only via the effect on parallelism, but also directly because judges have more cases to complete before starting to work on those assigned in the current quarter. Non-reported evidence based $\alpha_{i,q-1}$ as an instrument confirms this intuition.

current quarter. All estimates are statistically significant[20] and the signs correspond to the predictions of the theory. In particular, more task juggling measured by a larger number of new opened trials increases the average duration of all cases assigned during the current quarter. Similarly positive is the effect of a larger assigned workload in the quarter, while a greater standardized effort reduces duration and the implicit time trend is positive.

These estimates, however, are potentially inconsistent for the causal effect of interest. Column 2 reports Instrumental Variable (IV) estimates obtained using the instruments described above in Section 5.1. The effects of the confounded variables $\nu_{i,q}$ and $\left(\frac{e}{S}\right)_{i,q}$ are now larger and still statistically significant.[21] At the mean of the distribution of new opened cases (127)[22], ten fewer newly opened cases in a quarter (an 8% decrease of this indicator of task juggling) reduce the duration of assigned cases by 8.6 days (a 3% improvement, given a mean duration of 290 days). To put the size of this effect in the right perspective we can ask how many new hearings per quarter (for given difficulty of cases) the representative judge would have to hold in order to achieve the same reduction in the total duration of assigned trials. Given an estimate of -1.81 for the coefficient of $\left(\frac{e}{S}\right)_{i,q}$, 4.7 additional units of standardized effort per quarter (a 4% increase at the mean of this variable which is 128) would be needed to reduce the duration of assigned cases by the same amount of 8.6 days. In other words, at the mean, an 8% decrease of task juggling has the same effect as a 4% increase of effort. If the average number of hearings per case is $S = 3.2$, 4.7 units of standardized effort mean approximately 15 more hearings per quarters.

In the third column of Table 5 we report estimates that measure the degree of task juggling as the number of active cases $A_{i,q}$ on the desk of each judge at the end of a quarter. Also in this column all the estimates are statistically significant at standard levels[23] and the signs correspond to the predictions of the theory. Using the corresponding IV estimates of the fourth column to compare the size of the effects, ten fewer active cases in a quarter

[20] Since both the number of clusters and the number of time observations are small, here and in the other columns of the table we follow the suggestion of Angrist and Pischke (2009, pag. 296) and report the largest between classical and robust standard errors (which are the latter in all cases). We have also computed standard errors clustered by judge, clustered by judge-quarter, and HAC Newey-West standard errors that are respectively equal to 0.08, 0.06 and 0.08 for the effect of $\nu_{i,q}$, i.e. the main effect of interest in this column. These alternative standard errors are similar to our preferred estimate and do not change the statistical significance of the estimated coefficients.

[21] The table reports robust standard errors that are the largest between classical and robust (see footnote 20). The alternative standard errors clustered by judge, clustered by judge-quarter, and HAC Newey-West standard errors are respectively equal to 0.39, 0.34 and 0.37 for the effect of $\nu_{i,q}$.

[22] See the descriptive statistics in Table 4

[23] The table reports robust standard errors that are the largest between classical and robust standard (see footnote 20). The alternative standard errors clustered by judge, clustered by judge-quarter, and HAC Newey-West standard errors are respectively equal to 0.08, 0.09 and 0.07 for the effect of $A_{i,q}$.

(approximately a 5% decrease of this indicator of task juggling, at the mean of 210 active cases per quarter) reduce the duration of assigned cases by 6.2 days (a 2% improvement).[24] To achieve the same effect with more standardized effort per quarter the representative judge would have to increase it by 5.3 units. So, in this case, a 5% decrease of task juggling has the same effect as a 4% increase in standardized effort. If the average number of hearings per case is $S = 3.2$, 5.3 units of standardized effort mean approximately 17 more hearings per quarters.

The alphabetical procedure that assigns cases to judges, described in Section 2, ensures that our instruments are randomly assigned, but if the time unit of observation were sufficiently long, the same random assignment scheme would imply that our instruments should necessarily be weak. The tests described in Table 3 show that using quarters as the time unit of observation allows us to maintain a sufficient variability in the assignment process. This variability ensures that our instruments are sufficiently strong not to jeopardize our interpretation of the IV estimates of equation 5, as shown in the first stage estimates of Table 6. In columns 2 and 3 of this Table we report Cragg-Donald Wald F statistics. These test statistics are above the critical values computed in Stock and Yogo (2005), Table 5.2, that imply an IV bias equal to at most 10% of the OLS bias for the specification in which $P_{i,q} = \nu_{i,q}$ (critical value: 13.43) and to at most 15 % for the specification in which $P_{i,q} = A_{i,q}$ (critical value: 8.18). Note also that the three instruments have different effects on the two endogenous variables. As expected (see Section 5.1), in the first stage regression for standardized effort, $\frac{e}{S}$, the fraction of new urgent cases and the fraction of new difficult cases per quarter are estimated to have effects that are significant and with the expected sign, while the fraction of recently assigned cases beyond the "60 days" threshold is not statistically significant. In the first stage regressions of the measures of parallelism ν and A, instead, the "60 days" rule originates the most powerful instrument.

Thus, the empirical evidence confirms the theoretical predictions concerning the inefficiency of task juggling. Judges who are induced to juggle more tasks, i.e. to work according to a more parallel schedule because of the "60 days" rule, require more time to complete the cases assigned to them. The estimated causal effect is not only statistically significant but also quantitatively important in comparison to the causal effect of exerting more standardized effort in terms of more hearings per quarters or fewer hearings to close a case.

[24] The table reports robust standard errors that are the largest between classical and robust (see footnote 20). The alternative standard errors clustered by judge, clustered by judge-quarter, and HAC Newey-West standard errors are respectively equal to 0.30, 0.30 and 0.30 for the effect of $A_{i,q}$.

6 Switching costs

There is a large management literature surveyed by Mark *et. al.* (2008) suggesting the existence of a *disruption cost* of interruptions, measurable in terms of additional time to reorient back to an interrupted task after the interruption is handled. In this section we extend the previous analysis to study this collateral effect of task juggling and assess its relevance.

A full theoretical treatment of switching costs is developed in Appendix 9.3.3; here we limit ourselves to a heuristic analysis. We conceptualize these costs as decreasing the amount of "effective effort" applied to the project. If, for example, the judge needs to spend the morning reviewing a case he forgot all about, then a day spent working on that case corresponds to an "effective" effort of just the afternoon. Note that in the data we measure effective effort (observed number of hearings completed by the judge per unit of time).

Let us define "effective effort" as

$$\widetilde{e} = e/\left(1 + \phi P\right) \tag{7}$$

The parameter $\phi \geq 0$ captures the size of the switching cost. Note that effective effort is lower than "nominal effort" if and only if $\phi > 0$. Thus, $\phi > 0$ captures the case of switching costs (forgetful judge), while $\phi = 0$ implies no such cost. In modeling effective effort, we choose to scale nominal effort e by a factor that is decreasing in P because we want to capture the idea that when task juggling, measured by P, gets larger and the judge takes longer between two successive hearings of the same case, then switching costs are larger and so less effective work gets done.[25].

Since in the data we measure effort with the observed number of hearings completed by the judge per unit of time, what we observe is effective effort $\tilde{e}$, not nominal effort e. As a result, in the presence of switching costs, the equation 5 that we have estimated in the previous Section 5 should be written, more correctly, as

$$D_{i,q} = \beta_0 + \beta_1 \alpha_{i,q} + \beta_2 \left(\frac{\tilde{e}}{S}\right)_{i,q} + \beta_3 P_{i,q} + \beta_4 q + \delta_i + \epsilon_{i,q} \tag{8}$$

Nevertheless, for given value of ϕ, all the comparative statics laid out in Proposition 1 and 2 carry through unchanged. And specifically β_3 measures the effect of task juggling net of the effect of switching costs. However, the estimates of this equation, do not say anything on the parameter ϕ, which is included in our observed measure $\tilde{e}$ of effective effort.

[25]This specification is more consistent with switching costs being due to "lost memory" of a case rather than, say, to a retooling of some machine. In the latter case the switching costs might be fixed, independent of P.

To learn about the existence and relevance of switching costs we then proceed in the following way. First, we estimate the regression

$$\left(\frac{\tilde{e}}{S}\right)_{i,q} = \rho_0 + \rho_1 \alpha_{i,q} + \rho_3 P_{i,q} + \rho_4 q + \delta_i + \epsilon_{i,q} \tag{9}$$

where the coefficient ρ_3 estimates the effect of an increase of task juggling on effective effort $\tilde{e}$. The sign of this coefficient is informative about the possible presence of two separate effects. First, increasing the number of open cases may lead the worker to worker longer hours (increase in $\tilde{e}$ through an increase in e); second, as the number of active cases increases, switching costs may decrease effective effort $\tilde{e}$, for given number of hours worked (decrease in $\tilde{e}$ through an increase in ϕ).

Results of this estimation are reported in the first panel of Table 7. In the IV columns 2 and 4, the measures of task juggling $P_{i,q}$ are instrumented with the same instrument (6) used for equation (5). The coefficient ρ_3 is estimated to be positive but it is statistically significant only in the OLS regression in which $P_{i,q} = \nu_{i,q}$. So there is some evidence that more task juggling for exogenous reasons induces more effective effort but this evidence is weak. This lack of significance could depend on the fact that there are no switching costs but the judge *does not increase* her nominal effort to compensate for the exogenously induced greater task juggling. Or, alternatively, that the judge *decreases* nominal effort, but opening new trials involves switching costs.

In order to disentangle these combinations of effects we implement the strategy suggested by the following proposition.

Proposition 3. *Suppose the judge has a finite number of cases and adopts a partial rotation. Suppose that we take a specific schedule and change it by anticipating the opening date of a trial and all the following trials.*

a) Suppose there are no switching costs and nominal effort is constant. Then the duration of the last case to be completed is unchanged.

b) Suppose there are no switching costs and nominal effort increases with task juggling. Then the duration of the last case to be completed increases.

c) Suppose there are switching costs, and nominal effort is constant. Thus effective effort reacts to task juggling only because of switching costs. Then the last case to be completed is completed later.

Proof: Part a is proved in Appendix 9.3.2. Parts b c follow directly from Propositions 5 and **??** in Appendix 9.3.3. ■

The full theoretical framework behind this proposition is developed in Appendix 9.2. Here we simply give an intuitive account. Part a) simply reflects the fact that, in a finite set of cases, the last case to be completed requires all the steps of all the cases to have been carried out; the order in which they are carried out does not matter for the duration of the last case (provided the number of steps per quarter, i.e., effort, is constant). Part b) reflects the idea that, to the extent that congestion (working on many cases at the same time) reduces the "effective hours" that the judge can put in per quarter (net of the time spent brushing up on old cases), the effect of this "effective slowdown" accumulates, resulting in greater duration for all cases but especially for those which have suffered from the greatest number of "slowdowns," which are of course the last cases. Part c) is simply the obverse of part b).

Before taking this proposition to data, a caveat is in order. Proposition 3 focuses on the last case to be completed. When we go to the data we identify this last case with a proxy for last case assigned in quarter q.[26] This is because these cases are the last cases in the "shock" phase $\nu + \Delta$. However, these cases are not literally the last to be completed by the judge: in the framework of Section 4 the judge works on an infinite stream of cases. The last cases assigned in quarter q are merely the last cases to be completed which were assigned in the "shock phase." With this caveat well in mind, we now proceed.

In order to take Proposition 3 to the data, we estimate the equation

$$L_{i,q} = \lambda_0 + \lambda_1 \alpha_{i,q} + \lambda_3 P_{i,q} + \lambda_4 q + \delta_i + \Lambda_{i,q} \tag{10}$$

in which we proxy the duration of the longest case assigned to judge i in quarter q with the top quantiles of the duration distribution. Note that in this equation we do not control for standardized effective effort, since we posit that the effect of time use on durations is channeled through effective effort.

Results are reported in Panel B of Table 7. These quantile and quantile instrumental variables estimates have been obtained by implementing the Least Absolute Deviations estimator (QREG) and the Chernuzkov and Hansen (2008) Instrumental Variables Estimator (IVQREG).[27]

Focusing on the IV estimates, the effects on the 90th and 95th quantiles are respectively positive and negative but small in absolute size and largely insignificant. In the light of

[26] Specifically the 90th and 95th quantiles of the duration distribution of each quarter.

[27] We are grateful to Do Wan Kwak who shared with us his Stata code that implements the Chernuzkov and Hansen (2008) estimator. As a robustness check, we have also considered a special case of the Amemiya (1982) Least Absolute Deviations (2SLAD) estimator discussed in Chesher (2003), that provides identification conditions for quantile endogenous regressions. Estimates obtained with this method are similar to those reported in Table 7 and are available from the authors on request.

Proposition 3 these results suggest that the complementary effects of task juggling related to switching costs are probably small even if the mechanical effect on average duration, described by our model and estimated in Section 5 remains quantitatively and statistically significant.

7 Why do judges juggle tasks?

What leads our judges to operate in an apparently inefficient way? We believe that one reason is that scheduling/planning is mentally costly.

To understand the source of this cost, it helps to see the mechanics that create an inefficient schedule. The key source of inefficiency, in our case, is that judges tend to schedule a single case's hearings over time, rather than all at once. That is, a judge will wait to schedule the second hearing until after the first hearing has been held; will schedule the third hearing only after the second has been held; etc. This is convenient because it does not require the judge to guess how many hearings will be required for each case and how soon each can be held. The trouble with this process is that, if the assignment rate is too large, the judge will find that by the time he gets around to holding hearing h, the schedule is filled with extraneous hearings, including of newly assigned cases, so that the "first available" date for hearing $h+1$ is far in the future, farther than the technical time that the parties and the judge need to prepare for hearing $h+1$. So scheduling with this algorithm leads to hearings of a given case being spaced far apart in time. The time in between hearing h and $h+1$ is filled with hearings from other cases. This scheduling corresponds, on a massive scale, to the task juggling of Example 1 in the introduction.

Not all judges behave in this way, as we have shown in Section 3 and specifically in Figure 1. Moreover, as already mentioned in Section 5.1, some judges were even willing to risk being reprimanded because too many of their first hearings took place far beyond the "60 days" threshold (see Figure 6). And they were acquitted precisely because they defended themselves by showing that, in working sequentially, they achieved lower average durations than their colleagues.

Nevertheless, for many other judges task juggling seems convenient: it eliminates the need for the judge to forecast the pace of development of a case, and may help reduce the need for rescheduling. It is the avoidance of this planning effort, we believe, which leads to task juggling. Some judges may be better able to forecast, or more willing to incur these costs; other judges may be less able or willing. This heterogeneity, we suspect, explains at least some of the heterogeneity in scheduling behavior for our judges.

If mental costs are what is holding back "proper" scheduling, then the solution might be to give judges sufficiently sharp incentives based on productivity. Once the judge's objectives are tightly aligned with the principal's, then the judge will overcome her mental costs and schedule properly. The argument is sound, we believe, but it ignores a key problem: if the worker is steeply rewarded for output, then she may have a tendency to focus on easy/quick projects at the expense of complex ones. This may be problematic for the principal. For example, in Italy there is a concern that an excessive focus on productivity may lead some judges to not work on complex trials. We discuss this issue in our companion paper Coviello et al. (2011).

In our interactions with judges, other considerations emerged which lead some judges to forgo "proper" scheduling. When judges are overloaded, proper scheduling requires putting newly assigned cases "on ice" during an initial period of inactivity. In tribunals which are very overloaded, as is typical in the South of Italy, this initial period of inactivity might last several years. Some judges find it difficult to tell the parties that activity on their case will be delayed for such a long time, even if the judges know that the trial would in fact conclude earlier. Some judges raised a related concern for the "inequity" in the treatment of cases assigned at different points in time.

Another reason why workers (not necessarily our judges) juggle tasks is the "interdependent workplace." The Time Use literature defines the "interdependent workplace" as an environment in which other workers (co-workers, superiors, clients) can and do ask/demand immediate attention to projects which are relevant for them. Sometimes this results in a barrage of requests which may distract the worker from her more urgent tasks and lead the worker to juggle tasks. We do not think this is a relevant issues for our judges, but we analyze the interdependent workplace in our companion paper Coviello et al. (2011).

8 Conclusions

We presented empirical evidence in favor of the theoretical hypothesis that individual work scheduling and time use have significant effects on the speed at which workers can complete assigned jobs. We test this prediction on a sample of Italian judges and show that those who are exogenously induced to juggle more trials take more time to complete similar portfolios of cases.

This effect is a measure of the inefficiency of task juggling, which is largely practiced by workers in general and by Italian judges in particular. Such inefficiency appears puzzling given the assumption, common to most of academic economics, that workers operate on their

productivity frontier and avoid un-nececssary inefficiencies.

In order to identify the inefficiency of tasks juggling we construct time-varying instruments based on the sample realization of the lottery that allocates cases to each judge. This lottery is used in combination with the procedural rule prescribing that judges should hold the first hearing of a case no later than 60 days from filing. In this way exogenous increases in the number of assigned cases generate pressure for more task juggling around and after 60 days from filing. The effect that we measure is not only statistically significant but also quantitatively important: an exogenously induced 8% increase of task juggling would need to be compensated by a 4% increase of effort in order to avoid an increse in the average duration of trials. Our results also suggest that the negative effect of increased task juggling is partly offset by an endogenous increase in effort on the part of judges (more hearings). Switching costs, on the other hand, appear to be less significant in our setting. We believe our results provide the first empirical estimates of the impact of time allocation on productivity.

We have also considered the effect of task juggling on the measure of output quality at our disposal (percent of appeals to higher court), finding no adverse effect. For these judges, therefore, we do not detect a trade-off between quantity and quality. Obviously, this conclusion may vary when looking at different types of workers.

Our results have been derived for the specific setting of Italian judges, but the message of our paper concerning the effect of task juggling on the speed of job completion is more general because it applies to all those situations in which more output is required, but labor or capital cannot be increased at least in the short run. We view the analysis in this paper and its companion (Coviello et al. 2011) as a first step into the theoretical and empirical analysis of how work scheduling affects the individual production function.

9 Online Appendix

9.1 Derivation of an Equation for the Duration of Trials in a Stable Rotation

Let us start with some notation. For each quarter q, denote by α the number of cases assigned to the judge in that quarter, let ν denote the rate at which cases are opened in that quarter, let e denote the effort (number of tasks accomplished) in that quarter. Finally, let A denote the number of cases actively being worked on in a quarter. None of these quantities is indexed by q because in a stable rotation they will all be constant over time.[28]

Our task is to determine the ν that is compatible with the stable rotation, given the judge's effort e and the number of tasks S required to dispose a case. As there are A active cases at the beginning of a quarter, and since every time a case closes another one opens, at any instant within a quarter there are exactly A open cases. If we link any case that closes to the one that opens right after it closes, we have exactly A "links" in each quarter. Due to the procedure of rotation on the open, the judge must accomplish an equal number of tasks for each link. Since by assumption e tasks are accomplished in total in each quarter, it follows that exactly $\frac{e}{A}$ steps must accomplished for each link. This implies that, at the end of the quarter, those cases are completed which, at the beginning of the quarter, had less than $\frac{e}{A}$ steps remaining. How many are those cases? To find out, observe that since we are positing the same rate ν of input and output in every quarter, at any point in time there must be an equal number of cases which are x steps away from completion, regardless of x. For example, at the beginning of a quarter there are exactly as many cases needing 1 step to dispose (i.e., are almost done) as there are needing S steps (i.e., are just beginning). Given this observation, we can compute how many cases have less than $\frac{e}{A}$ steps remaining at the beginning of a quarter: they are a fraction $\left(\frac{e}{A}\right)/S$ of the total number A of cases open at the beginning of the quarter. Therefore, in steady state the number of cases adjudicated in a quarter is given by

$$\frac{\frac{e}{A}}{S}A = \frac{e}{S}.$$

In other words, a steady state requires that cases be opened at the rate of $\frac{e}{S}$ per quarter. If cases are opened at this rate, then exactly $\frac{e}{S}$ cases are adjudicated in each quarter.[29]

Now let us work out the duration of a case. In a steady state cases are completed at the rate of ν per quarter. Then, given that α cases are assigned per quarter, a case assigned in quarter q finds

$$A_0 + \alpha q - vq$$

unfinished cases in front of it.[30] The duration D_q of a case is essentially the time it takes to adjudicate the unfinished cases that precede it. Given a completion rate ν, this duration is

$$D_q = \frac{A_0 + \alpha q - \nu q}{\nu}.$$

Plugging $\nu = e/S$ this into this equation yields equation (2).

[28] In steady state the judge works on A active cases in all quarters, including $q = 0$. One way to think about the presence of A at the beginning of a judge's tenure is that every incoming judge inherits the case load of the outgoing judge which he replaces.

[29] If more than e/S cases are opened in a quarter then the rate at which cases are adjudicated falls below e/S. We will show this in the next section.

[30] The presence of the term A_0 reflects the fact that we are assuming that in every period starting from $q = 0$, there are A cases actively being worked on.

9.2 Proof of Proposition 2

Fix arbitrarily e and S. Let $x \in [0, S]$ be a continuous variable denoting the state of completion of a case. Time is discrete and, in each period $1, 2, ...$ Three things happen sequentially. First, a mass ν_t of new cases is introduced at $x = S$. Then A_t is computed, which is the mass of all cases with $x \in (0, S]$. Then the x of all cases is decreased by e/A_t (this formula embodies the assumption of rotation on the open). The cases whose x falls at or below zero are called complete in period t. We denote their mass by ω_t.

We start from a careful description of stable rotation. Let us define a stable rotation as a case in which $\nu_t = \nu = e/S = \omega$. This means that in each period exactly e/S cases are assigned and opened, and exactly the same number are completed. Let us, for convenience, assume that A is a multiple of ν, or equivalently that S is a multiple of e/A. This assumption guarantees that, as we look at the docket in a stable rotation, it is composed of masses of cases that are equally spaced between 0 and S. This implies that, in a stable rotation, cases which get done do so by touching, not crossing, zero.

Let us now perturb a stable rotation at times t_0 and $t_0 + 1$ as follows. Let $\nu_{t_0} = \nu + \Delta$, $\nu_{t_0+1} = \nu - \Delta$, and otherwise $\nu_t = \nu$. We want to study the effect of this perturbation on the completion time of a case assigned at $t_0 + 1$.

Proposition 4. *Assume a judge employes a rotation on the open. Then case assigned at time $t_0 + 1$ under the perturbed schedule takes longer to complete (longer duration) than a case assigned at the same time under a stable rotation schedule.*

Proof: In a stable rotation the duration of a case is given by the time it takes for it to hit zero. Since such a case must cover a distance from S to 0, and in a stable rotation cases move to the left by exactly e/A in each period, it takes $S/\left(\frac{e}{A}\right)$ periods to travel that distance.

To compute durations under the perturbed rotation we must carefully trace the evolution of the system following t_0.

Time t_0 Since $A_{t_0} = A + \Delta$, movement to the left at time t_0 equals $e/(A + \Delta)$. This is less than the movement that, in a stable rotation, was barely sufficient to clear a case. Hence in period t_0 no cases are cleared

Time $t_0 + 1$ Before the injection of ν_{t_0+1} we have a mass of cases equal to $A + \Delta$. After $\nu_{t_0+1} = \nu - \Delta$ is added we have $A_{t_0+1} = A + \nu$. This means that the translation to the left in period $t_0 + 1$ equals $e/(A + \nu)$. This is less than $\frac{e}{A}$, which is the distance between the first and second most complete batch of cases at the beginning of period $t_0 + 1$ (these cases were inputed under a stable rotation). This means that at the end of period $t_0 + 1$ we cannot have cleared more than ν cases and the end-of-period mass of active cases is at least A.

Time $t_0 + 2$ Before the injection of ν_{t_0+2} we have a mass of cases equal to A. After $\nu_{t_0+2} = \nu$ is added we have $A_{t_0+2} = A + \nu$. This means that the translation to the left in period $t_0 + 2$ equals $e/(A + \nu)$. This is less than $\frac{e}{A}$, which is the distance between the first and second most complete batch of cases at the beginning of period $t_0 + 2$ (these cases were inputed under a stable rotation). This means that at the end of period $t_0 + 2$ we cannot have cleared more than ν cases and the end-of-period mass of active cases is at least A.

Time $t_0 + ...$ Same as time $t_0 + 2$, until that time period in which cases assigned at time t_0 clear. (In that period a larger mass than ν clears, and so one cannot conclude that "the end-of-period mass of active cases is at least A."

Let us consider now the duration of a case assigned at time t_0+1. In order for it to clear there can only be two cases. First case is that these cases clear in the same period as those assigned at time t_0. In this case in each period of their "active life" the t_0+1 cases move by translations of size no longer than $e/(A+\nu)$. Since these cases need to travel a distance equal to S to complete, this means that it takes these cases no less than $S/\frac{e}{A+\nu}$ periods to complete (and, in the likely case that this number is not an integer, it takes exactly $int\left\{S/\frac{e}{A+\nu}\right\}+1$ periods to complete the cases assigned at time t_0+1.) Therefore, a case assigned at time t_0+1 under the perturbed schedule takes longer to complete than a case assigned at the same time under a stable rotation schedule.

The second case we need to cover is that in which the cases assigned at time t_0+1 complete at a later period than those assigned at time t_0. In this case in the last period of their "active life" the t_0+1 cases move by translations of a size which might be longer than $e/(A+\nu)$. Our computation in this case is indirect. At the beginning of period t_0+1 the cases assigned at time t_0 lie at $S-[e/(A+\Delta)]$. From period t_0+1 onward these cases translate to the left by an amount no larger than $e/(A+\nu)$ per period. Therefore these cases will take, between t_0+1 and the time when they are done, a number of periods no smaller than

$$\frac{S-[e/(A+\Delta)]}{e/(A+\nu)}$$

(and if this quantity is not an integer, then in a number of periods equal to its integer part plus 1, which is greater than this quantity.) Since we are in the scenario where the cases assigned at time t_0+1 clear in a period subsequent to those assigned at t_0, the former cases (which are our focus) take at least

$$\frac{S-[e/(A+\Delta)]}{e/(A+\nu)}+1$$

periods to complete (and, if this quantity is not an integer, then in a number of periods equal to its integer part plus 1, which is greater than this quantity.) Now observe that

$$\begin{aligned}
& \frac{S-[e/(A+\Delta)]}{e/(A+\nu)}+1 \\
= \;& \frac{S}{e}(A+\nu)-\frac{(A+\nu)}{(A+\Delta)}+1 \\
= \;& \frac{S}{e}(A+\nu)-\frac{\nu-\Delta}{(A+\Delta)} \\
> \;& \frac{S}{e}(A+\nu)-1 \\
= \;& \frac{S}{e}A+\frac{S}{e}\nu-1=\frac{S}{e}A,
\end{aligned}$$

where the last equality follows from the fact that in a stable rotation $\nu=e/S$. Therefore, a case assigned at time t_0+1 under the perturbed schedule takes longer to complete than a case assigned at the same time under a stable rotation schedule. ■

Proposition 4 proves the first part of Proposition 2. Now we turn to the second part of Proposition 2. Under the assumed work scheduling practice in that second part, the judge opens ν_t cases per quarter but only works on a subset of the open cases. The not-yet-worked on set of open cases lie at S, waiting for the cases which precede them to be worked on. By assumption, the set of cases the judge works in each period is independent of the rate ν_t at which cases are opened. This implies that the output rate, and even the identity of the

cases closed in each period, does not depend on the rate at which cases are opened. The only two determinants of the duration of a case is the date in which it was assigned, and how many cases arrived before it. In our experiment, we keep constant the order in which cases are assigned, and we look at the duration of cases which are assigned at the same time q. Therefore, the duration of these cases does not depend on ν_q.

9.3 Proof of Proposition 3

In this section we develop a theory of task juggling where task juggling is understood as anticipating the opening of some cases *and all the cases that follow them*. This is exactly the perturbation presented in Proposition 2 part b. The main goal in this section is to prove Proposition 3 which, in turn, will prove Proposition 2 part b.

Of note, the main result in this section (Proposition 3) applies to the case in which the judge deals with a finite–not an infinite–number of cases. We emphasize this aspect of our analysis in the main text where it matters.

9.3.1 Setup and Definitions

Time is indexed by quarters q, starting with $q = 1$, the first quarter in which the judge operates, and possibly going to infinity.

A judge confronts C cases, where a case is indexed by c. We allow C to be infinite. Each case is made up of S distinct steps, or tasks, each of which takes 1 unit of time to accomplish. The s-th step of case c is denoted by c_s. Case c is said to be *completed* when its last step c_S has been accomplished.

Cases begin by being assigned to a judge. Cases may not all be assigned at once; rather, they may be assigned progressively over time. As a matter of convention, we stipulate that cases with lower c arrive earlier. We denote with α_q the number of cases assigned in quarter q. Each case is worked on progressively through several quarters, and in each quarter the judge works on several cases. The number of tasks (steps of possibly different cases) accomplished in quarter q is denoted by e_q. We interpret e_q as capturing the judge's effort in quarter q.

All cases assigned in quarter q are assumed to take S_q tasks to dispose. In our empirical analysis, S_q is measured as the average number of hearings it takes to adjudicate a case. Clearly, as we said in Section 3.3, this measure reflects the inherent complexity of the cases assigned to the judge. Moreover, to the extent that S_q varies systematically across judges even though workloads are of similar complexity, S_q also reflects some kind of individual ability of the judge, the ability to adjudicate cases with fewer hearings. In general we will interpret S_q as a measure of both complexity of cases and individual ability of the judge. When comparing identical portfolios of cases assigned to different judges, instead, it will measure their ability.

The duration of case c is the number of quarters that elapse between the time the case is assigned and the time it is completed. We denote the duration of a case assigned in quarter q as D_q.[31]

We now discuss the ways in which the judge allocates his effort across cases and through time. To this end we introduce the notion of *work schedule*. A work schedule simply captures the order in which the judge accomplishes tasks related to different cases. We will define two polar opposite work schedules, the *sequential* and the *full rotation* schedules. We then define a third, more general type of work schedule, which we call *rotation on the open*. As we will show, both the sequential and the full rotation schedule are special cases of rotation on the open.

[31] Even within our stylized models it is possible that the cases assigned at the beginning of quarter q may be disposed earlier than those assigned at the end of quarter q. In this case one might want to consider more complicated measures of duration, such as the average duration of cases assigned in a quarter. To sidestep this inconvenience, we define D_q as the duration of the first case assigned in a quarter.

For ease of exposition in this subsection we assume that all the cases have been assigned in the first quarter.

Definition 2. *A **work schedule** is a complete strict order $\prec$ on the set of all tasks such that*
a) $c_s \prec c_{s'}$ if $s < s'$.
b) $c_1 \prec c'_1$ if $c < c'$.

The first condition says that the steps of case c have to be performed sequentially, from first to last. This requirement does not mean that the steps have to be performed consecutively—the judge can alternate between steps of different cases. The second condition says that a case with a higher index cannot be started before any case with a lower index.

We now define three different work schedules.

Definition 3. *The **sequential** work schedule is the work schedule in which the ordering $c_s \prec c'_{s'} \prec c_{s+1}$ does not arise for any $c_s, c'_{s'}$.*

*The **full rotation** work schedule is one in which, between every two steps of a given case, there is at least one task of every other case. Formally, given c_s, c_{s+1}, for any $c' \neq c$ there is some s' such that $c_s \prec c'_{s'} \prec c_{s+1}$.*

*A **rotation on the open** is a work schedule in which if $c'_1 \prec c_s \prec c'_S$ then there is some s' such that $c_s \prec c'_{s'} \prec c_{s+1}$.*

The sequential work schedule is that in which cases are worked on sequentially: first all the steps relating to the first case are accomplished, then all the steps relating to the second case, etc. The polar opposite of a sequential work schedule is the full rotation one, in which, within each step, cases are worked on according to their arrival order. In Lemma 1 we show that, in a full rotation, c_s must immediately be followed by $(c+1)_s$ and C_s must immediately be followed by 1_{s+1}. A rotation on the open is a process that works just as a full rotation does, except that instead of rotating on all cases, the rotation on the open does not touch cases that have not been started yet. The condition $c'_1 \prec c_s \prec c'_S$ identifies those cases c' that were open at the time step c_s was accomplished.

Example 1. *Let there be three cases each requiring two steps, so that $C = 3$ and $S = 2$. The sequential work schedule is*

$$1_1 \prec 1_2 \prec 2_1 \prec 2_2 \prec 3_1 \prec 3_2.$$

The full rotation work schedule is

$$1_1 \prec 2_1 \prec 3_1 \prec 1_2 \prec 2_2 \prec 3_2.$$

Now let there be three cases each requiring three steps, so that $C = 3$ and $S = 3$. The following schedule is a rotation on the open.

$$1_1 \prec 2_1 \prec 1_2 \prec 2_2 \prec 1_3 \prec 3_1 \prec 2_3 \prec 3_2 \prec 3_3.$$

In the first five positions of the schedule only cases 1 and 2 are open, and so the definition of rotation requires the schedule to alternate between the steps of cases 1 and 2. In the sixth position case 3 gets opened. The definition then requires that 2_3 follow, because the alternative (3_2) would violate the definition (set $c = 3, c' = 2$.) The fact that 3_2 and 3_3 are adjacent in the order does not violate the definition because only case 3 is open by the time the order gets to the last two tasks.

Let us contrast the full rotation and sequential work schedule. In the full rotation schedule cases are started as early as possible, and they are completed late in the order of the work schedule. Consequently, at any given point in time there is a large mass of cases being simultaneously worked on. In contrast, a sequential work schedule causes the start of a new case to be postponed as late as possible, and the completion of cases happens evenly throughout the unfolding of the work schedule. As a result, in a sequential work schedule the minimum possible number of cases is simultaneously being worked on at any point in time. In this sense, we can say that a full rotation is the polar opposite of a sequential work schedule.

The rotation on the open is a general family of work schedules which subsumes as special cases the full rotation and the sequential. This family is parameterized by the position in the work schedule in which cases are opened. That is to say, a rotation on the open can take different forms depending on how early in the work schedule the cases are opened. If, for example, all cases are opened as early as possible, and thus the first C steps in the ordering are $1_1, 2_1, ..., C_1$, then a rotation on the open becomes identical to a full rotation. If, instead, new cases are opened at the slowest possible pace, that is, one every S steps, then there is only one case open at the any one time and so the rotation on the open becomes a sequential work schedule.

Lemma 1. *In a full rotation, c_s is immediately followed by $(c+1)_s$ and C_s is immediately followed by 1_{s+1}.*

Proof. The first element of the order must by definition be 1_1. The definition of full rotation implies that, between the first and second step of the first case, 1_1 and 1_2, there must be tasks $2_1, 3_1, ..., C_1$. By Definition 2 b, these tasks must be ordered as $2_1 \prec 3_1 \prec ... \prec C_1$. This shows that c_1 is immediately followed by $(c+1)_1$. Now, we claim that only these tasks can lie between 1_1 and 1_2. Suppose by contradiction that there was some c_2 in between 1_1 and 1_2. Then we would have $1_1 \prec c_1 \prec c_2 \prec 1_2$, which violates the definition of full rotation (set $c' = 1$). Next, let us show that C_1 is immediately followed by 1_2. Again, suppose not: then it would be immediately followed by some $c_2 \neq 1_2$. In this case we have a contradiction of Definition 2 b (set $c' = 1$). Reasoning by induction establishes the full statement of the Lemma. □

The following Lemma will be used in the proof of Proposition 3.

Lemma 2. *In a rotation on the open, if $c_s \prec c'_1 \prec c_{s+1}$ then $c_{k+s} \prec c'_{k+1} \prec c_{k+s+1} \prec c'_{k+2}$ for $k = 0, ..., S-(s+1)$.*

Proof. Let $\widehat{k}$ denote the lowest k at which there is a violation of the statement of the lemma. First, let us rule out $\widehat{k} = 0$. If $\widehat{k} = 0$ the only work schedule that violates the statement takes the form $c_s \prec c'_1 \prec c'_2 \prec c_{s+1}$. But this contradicts the definition of rotation on the open (just switch c and c'). Therefore it must be $c_s \prec c'_1 \prec c_{s+1} \prec c'_2$. Suppose then that $\widehat{k} = 1$. This means that $c_{s+1} \prec c'_2 \prec c_{s+2} \prec c'_3$ is violated. There are only two work schedules which violate this. One is $c_{s+1} \prec c'_2 \prec c'_3 \prec c_{s+2}$, and this is not a rotation on the open (just switch c and c' in Definition 3). The other is $c_{s+1} \prec c_{s+2} \prec c'_2 \prec c'_3$, which violates Definition 3 (case c' is open while the judge executes steps c_{s+1} and c_{s+2}). Since both violations are not compatible with the definition of rotation on the open, it cannot be that $\widehat{k} = 1$. Reasoning by induction proves the lemma. □

This property says that, as soon as a case c' is started, its steps are accomplished in lockstep with the steps of all other cases already open, in the sense that the schedule will rotate among the steps of these cases in the same order. This does not mean, of course, that the interval between two tasks c_{k+s} and c'_k is always the same as k progresses That will depend on how many other cases are being opened and closed as the schedule unfolds.

9.3.2 Proof of Proposition 3 part a

To simplify the exposition we will maintain the assumption that all cases have been assigned in the first quarter. Our model then implies that all cases take $S_1 = S$ to complete. At the end of the section we will discuss what happens if cases are heterogeneous in the number of steps they take to complete.

Definition 4. *The **rank** $\rho(c_s)$ of task c_s is given by 1 plus the number of tasks which precede c_s in the ordering of the work schedule. The **opening rank** of case c is $\rho(c_1)$. The **completion rank** of case c is $\rho(c_S)$.*

Although the previous definition does not explicitly involve quarters, one may still associate $\rho(c_s)$ with the time period in which task c_s is performed. If $\rho(c_s)$ is small then we think of that task as being performed earlier. Thus, for example, we say that case c is completed *earlier* if $\rho(c_S)$ becomes smaller.

A main focus of our analysis is the early completion of cases. We want to show that, within the family of rotations on the open, anticipating the opening of cases tends to delay the completion of all cases. To this end, we need to be precise about what it means to anticipate the opening of cases.

Definition 5. *Take two rotations on the open denoted by $\prec$ and $\widetilde{\prec}$ with opening ranks given by $\rho(c_1)$ and $\widetilde{\rho}(c_1)$, respectively. We say that $\widetilde{\prec}$ **anticipates the opening** of case $\widehat{c}$ relative to $\prec$ if: (a) the work schedules $\prec$ and $\widetilde{\prec}$ coincide at ranks lower than $\widetilde{\rho}(\widehat{c}_1)$; and (b) $\widetilde{\rho}((\widehat{c}+k)_1) - \widetilde{\rho}(\widehat{c}_1) = \rho((\widehat{c}+k)_1) - \rho(\widehat{c}_1)$.*

This definition says that anticipating the opening of case $\widehat{c}$ means the following. Starting from a rotation on the open ρ, one decreases the opening ranks of all cases $\widehat{c}$ and higher by the same amount. In order to end up with a rotation on the open, this will require rearranging the ordering of tasks above $\widetilde{\rho}(\widehat{c}_1)$. Otherwise, the ordering of tasks below $\widetilde{\rho}(\widehat{c}_1)$ is left unchanged. Let's work through an example.

Example 2. *Consider the following two rotations on the open.*

$$1_1 \prec 2_1 \prec 1_2 \prec 2_2 \prec 1_3 \prec 2_3 \prec 3_1 \prec 4_1 \prec 3_2 \prec 4_2 \prec 5_1 \prec 3_3 \prec 4_3 \prec 5_2 \prec 5_3$$
$$1_1 \prec 2_1 \prec 3_1 \prec 4_1 \prec 1_2 \prec 2_2 \prec 5_1 \prec 3_2 \prec 4_2 \prec 1_3 \prec 2_3 \prec 5_2 \prec 3_3 \prec 4_3 \prec 5_3$$

In the second schedule the openings of case 3 and all following cases are anticipated by 4 periods, relative to the first schedule. Now let's look at the date of completion. Cases 1 through 4 are completed later in the second schedule than in the first, while case 5 is completed at the same time in the two schedules.

This example shows what it means to anticipate cases 3 and following. In the example the opening of case 3 is moved up to the place in the order where cases 1 and 2 get opened. The effect of this perturbation is to increase the "frequency" with which cases are opened early on in the order, and otherwise leave unchanged the "frequency" with which cases are opened (except for the end of the order, where fewer cases are opened because there are no more cases to open). The example also shows that the effect of such anticipation is to increase the completion rank of all cases.

Proof of Proposition 3.

Proof. First, observe that all cases $c < \widehat{c}$ will obviously last weakly longer, and be disposed no earlier, after the anticipation. Let us turn to the time at which cases $\widehat{c}, \widehat{c}+1, ...$ are disposed.

Consider now two cases c, c' with $\widehat{c} < c < c'$. The relative order in which tasks from c and c' are performed is fully determined once we know the index s that solves $c_s \prec c'_1 \prec c_{s+1}$. That is because, by Lemma 2, once two cases have been started they go in lockstep forever

after, meaning that the relative ordering of their tasks does not change even though the time that elapses between them changes as other cases are opened and closed. Now, if $\rho(c_1') - \rho(c_1) = k$ (which means that under the old schedule c' was opened k periods after c was opened) under the new schedule we still must have $\widetilde{\rho}(c_1') - \widetilde{\rho}(c_1) = k$. However, if there was some c_s being accomplished before c_1', that is $\rho(c_s) < \rho(c_1')$, it is not guaranteed that $\widetilde{\rho}(c_s) < \widetilde{\rho}(c_1')$. This is because there may be open cases at the time that c is started whose steps must be accomplished before c_2 is performed, and that may well push c_2 (or more generally c_s) until after c_1'. This means that $\widetilde{\rho}(c_s) > \widetilde{\rho}(c_1')$ if $\rho(c_s) > \rho(c_1')$, but the converse is not necessarily true. Now, once case $c' > c$ has been started, then c and c' are accomplished in lockstep forever after, meaning that the relative ordering of their tasks does not change even though the time that elapses between them changes as other cases are opened and closed. Therefore, by the time c_S is done, there are fewer steps of case c' left to accomplish relative to the initial schedule.

Now, set $c = C - 1$ and $c' = C$. There can be no tasks of cases of index smaller than c between c_S and c_S' because cases opened earlier are finished before cases opened later. Only steps of case c' can be left to accomplish. Then our result implies that $\widetilde{\rho}(C_S) - \widetilde{\rho}((C-1)_S) \leq \rho(C_S) - \rho((C-1)_S)$. Since $\widetilde{\rho}(C_S) = \rho(C_S)$, it follows that $\widetilde{\rho}((C-1)_S) \geq \rho((C-1)_S)$, that is, case $C-1$ is accomplished later due to the anticipation.

Now set $c = C - 2$. Setting $c' = C$ implies that there are fewer steps of case C to accomplish between $(C-2)_S$ and C_S. Setting $c' = C-1$ implies that there are fewer steps of case $C-1$ to accomplish between $(C-2)_S$ and $(C-1)_S$. Since only steps of cases C and $C-1$ can arise between $(C-2)_S$ and C_S, we have shown that there are fewer steps of any case that are performed between $(C-2)_S$ and C_S. Thus, $\widetilde{\rho}(C_S) - \widetilde{\rho}((C-2)_S) \leq \rho(C_S) - \rho((C-2)_S)$. Since $\widetilde{\rho}(C_S) = \rho(C_S)$, it follows that $\widetilde{\rho}((C-2)_S) \geq \rho((C-2)_S)$, that is, case $C-2$ is accomplished later due to the anticipation.

Reasoning analogously, one can show that any case $c' > \widehat{c}$ is disposed no earlier due to the anticipation. □

This proposition says that anticipating the opening of a case, and all that follow, imposes a negative externality on all other cases if the judge follows a rotation on the open. The intuition is simple. By opening a new case, the judge pulls resources away from cases which are closer to being completed i.e., all other cases given the First In First Out (FIFO) nature of a rotation on the open. Moreover, the newly opened case does not benefit from being opened earlier, in the sense that it will still have to wait that all other cases are completed before it too can be completed (again, this follows from the FIFO nature of the rotation on the open). Therefore, opening too many cases too early is Pareto-inferior.

This proposition also implies that all cases last longer in a full rotation schedule than in a sequential schedule. Indeed, a full rotation schedule is obtained starting from a sequential schedule and progressively anticipating the opening of all cases $2, ..., C$. More generally, the proposition implies that an efficient judge is one who opens cases at a slow rate and keeps few cases active at any given time.

9.3.3 Switching Costs, and Proof of Proposition 3 parts b and c

Until now we have identified the duration of a case c with $\rho(c_S)$, the number of steps completed before the last step of case c. This interpretation reflects the idea that all steps of all cases takes approximately the same amount of time each, and that the time elapsed before case c is completed is simply the sum of these times. We now want to enrich this model. We want to allow for the possibility of time costs which arise when a case is put down and then started again much later. These costs reflect the extra time it takes a judge to remember where he left off. If we capture this "wasted" time on task τ by some function

$W(\tau, \rho)$, then the duration of case c is given by

$$\rho(c_S) + \sum_{\tau \text{ s.t. } \rho(\tau) \leq \rho(c_S)} W(\tau, \rho).$$

Note that now the duration of case c depends not only the number of tasks completed before c_S, the last step of case c, but also on the "time wasted" on all those tasks. We refer to the second addend as to "time wasted up to the completion of case c."

To model setup costs a simple functional form will suffice:

$$\begin{aligned} W(c_1, \rho) &= \underline{W} \\ W(c_s, \rho) &= w(\rho(c_s) - \rho(c_{s-1})) \quad \text{for } s > 1. \end{aligned} \tag{11}$$

This formulation captures the idea that the first step of any case entails a setup cost $\underline{W} \geq 0$ (necessary waste of time due to the necessity of collecting and organizing basic information about the case). This setup cost does not depend on the judge's work practices ρ, and is, for simplicity, the same for all cases. Furthermore, the time wasted on each subsequent step of a case depends only the difference between the rank of that step, and the rank of the preceding step of the same case. We require that

$$w(\cdot) \text{ is a non-decreasing function.} \tag{12}$$

This means that if task c_s is accomplished soon in the order after the preceding step in the same case, then the time wasted time on task c_s is small. We also impose that

$$w(\cdot) \geq 0, \tag{13}$$

which means that time wasted cannot be negative. Finally, for simplicity, we normalize the minimum possible time wasted to zero:

$$w(1) = 0. \tag{14}$$

Thus, if two steps of the same case are accomplished sequentially, then there is zero time wasted on the second step.

Proposition 5. *Consider two rotations on the open ρ and $\widetilde{\rho}$ with such that $\widetilde{\rho}$ anticipates the opening of case $\widehat{c}$ relative to ρ. The time wasted up to the completion of each case increases as we switch from ρ to $\widetilde{\rho}$, and this effect is stronger for cases completed later.*

Proof. We need to show that the difference under $\widetilde{\rho}$ and under ρ in time wasted before the completion of any case κ is positive and nondecreasing in κ.

$$\begin{aligned} &\sum_{\rho(\tau) \leq \rho\left((\kappa+1)_S\right)} W(\tau, \rho) - \sum_{\rho(\tau) \leq \rho(\kappa_S)} W(\tau, \rho) \\ = &\sum_{\substack{c_s \text{ s.t. } \rho(\kappa_S) < \rho(c_s) \leq \rho\left((\kappa+1)_S\right) \\ s > 1}} w(\rho(c_s) - \rho(c_{s-1})) + \sum_{c_1 \text{ s.t. } \rho(\kappa_S) < \rho(c_1) \leq \rho\left((\kappa+1)_S\right)} \underline{W}. \end{aligned}$$

Now observe that the set $\{c_s \text{ s.t. } \rho(\kappa_S) < \rho(c_s)\}$ is the same as the set $\{c_s \text{ s.t. } c \geq \kappa + 1\}$ because in a stable rotation cases are finished in the order they are opened. Moreover, the set $\{c_s \text{ s.t. } \rho(c_s) \leq \rho((\kappa+1)_S)\}$ is the same as the set $\{c_s \text{ s.t. } c \geq \kappa + 1 \text{ and } \rho(c_1) < \rho((\kappa+1)_S)\}$ because if some case c has a step before the last step of case $\kappa + 1$, then that case must

have been started before case $\kappa+1$'s last step, and not yet finished (which, in a stable rotation is guaranteed by the condition that $c \geq \kappa+1$.). This means that the set $\{c_s \text{ s.t. } \rho(\kappa_S) < \rho(c_s) \leq \rho((\kappa+1)_S)\}$ is the same as the set $\{c_s \text{ s.t. } c \geq \kappa+1 \text{ and } \rho(c_1) < \rho((\kappa+1)_S)\}$. Therefore the previous expression equals

$$\begin{aligned}
&\sum_{\substack{c_s \text{ s.t. } c\geq\kappa+1 \\ \rho(c_1)<\rho\left((\kappa+1)_S\right) \\ s>1}} w\left(\rho\left(c_s\right)-\rho\left(c_{s-1}\right)\right)+\sum_{\substack{c_1 \text{ s.t. } c\geq\kappa+1 \\ \rho(c_1)<\rho\left((\kappa+1)_S\right)}} \underline{W} \\
\leq &\sum_{\substack{c_s \text{ s.t. } c\geq\kappa+1 \\ \widetilde{\rho}(c_1)<\widetilde{\rho}\left((\kappa+1)_S\right) \\ s>1}} w\left(\rho\left(c_s\right)-\rho\left(c_{s-1}\right)\right)+\sum_{\substack{c_1 \text{ s.t. } c\geq\kappa+1 \\ \widetilde{\rho}(c_1)<\widetilde{\rho}\left((\kappa+1)_S\right)}} \underline{W} \\
\leq &\sum_{\substack{c_s \text{ s.t. } c\geq\kappa+1 \\ \widetilde{\rho}(c_1)<\widetilde{\rho}\left((\kappa+1)_S\right) \\ s>1}} w\left(\widetilde{\rho}\left(c_s\right)-\widetilde{\rho}\left(c_{s-1}\right)\right)+\sum_{\substack{c_1 \text{ s.t. } c\geq\kappa+1 \\ \widetilde{\rho}(c_1)<\widetilde{\rho}\left((\kappa+1)_S\right)}} \underline{W} \\
= &\sum_{\widetilde{\rho}(\tau)\leq\widetilde{\rho}\left((\kappa+1)_S\right)} W\left(\tau,\widetilde{\rho}\right)-\sum_{\widetilde{\rho}(\tau)\leq\widetilde{\rho}(\kappa_S)} W\left(\tau,\widetilde{\rho}\right)
\end{aligned}$$

where the first inequality reflects the fact that $\widetilde{\rho}$ anticipates the first step of all cases following $\widehat{c}$, and so given any task $(\kappa+1)_S$ there will be more first steps in front of it under $\widetilde{\rho}$ than under ρ. The second inequality holds because, as we go from ρ to $\widetilde{\rho}$, the difference between two successive steps of any given case does not shrink, and it increases for some. Formally, this implies that for each $s>1$ we have $\rho(c_s)-\rho(c_{s-1}) \leq \widetilde{\rho}(c_s)-\widetilde{\rho}(c_{s-1})$.

Rearranging the two extremes in the chain of inequalities yields

$$\sum_{\widetilde{\rho}(\tau)\leq\widetilde{\rho}(\kappa_S)} W\left(\tau,\widetilde{\rho}\right)-\sum_{\rho(\tau)\leq\rho(\kappa_S)} W\left(\tau,\rho\right) \leq \sum_{\widetilde{\rho}(\tau)\leq\widetilde{\rho}\left((\kappa+1)_S\right)} W\left(\tau,\widetilde{\rho}\right)-\sum_{\rho(\tau)\leq\rho\left((\kappa+1)_S\right)} W\left(\tau,\rho\right) \quad (15)$$

The last equality shows that the difference under $\widetilde{\rho}$ and under ρ in time wasted before the completion of any case κ is nondecreasing in κ. To finish the proof, we need to show that this difference is positive for all κ. Since the difference is nondecreasing in κ, it suffices to show this for $\kappa=1$. In this case the left hand side of (15) capures the waste of time accumulated by the time the first case is completed. The time wasted up to 1_S (the last step of case 1) under ρ is

$$\begin{aligned}
\sum_{\rho(\tau)\leq\rho(1_S)} W\left(\tau,\rho\right) &= \sum_{\substack{c_s \text{ s.t. } \rho(c_1)\leq\rho(1_S) \\ s>1}} w\left(\rho\left(c_s\right)-\rho\left(c_{s-1}\right)\right)+\sum_{c_1 \text{ s.t. } \rho(c_1)\leq\rho(1_S)} \underline{W} \\
&\leq \sum_{\substack{c_s \text{ s.t. } \widetilde{\rho}(c_1)<\widetilde{\rho}(1_S) \\ s>1}} w\left(\rho\left(c_s\right)-\rho\left(c_{s-1}\right)\right)+\sum_{c_1 \text{ s.t. } \widetilde{\rho}(c_1)<\widetilde{\rho}(1_S)} \underline{W} \\
&\leq \sum_{\substack{c_s \text{ s.t. } \widetilde{\rho}(c_1)<\widetilde{\rho}(1_S) \\ s>1}} w\left(\widetilde{\rho}\left(c_s\right)-\widetilde{\rho}\left(c_{s-1}\right)\right)+\sum_{c_1 \text{ s.t. } \widetilde{\rho}(c_1)<\widetilde{\rho}(1_S)} \underline{W} \\
&= \sum_{\widetilde{\rho}(\tau)<\widetilde{\rho}(1_S)} W\left(\tau,\widetilde{\rho}\right).
\end{aligned}$$

The first inequality holds because for any given task (including 1_S), going from ρ to $\widetilde{\rho}$ does not decrease (and sometimes increases) the set of tasks completed before it. The second

inequality holds because the difference between two successive steps of any case does not shrink, and it increases for some, when we go from ρ to $\widetilde{\rho}$. The last expression is the time wasted up to 1_S under $\widetilde{\rho}$. □

This proposition shows that time wasted increases as we switch from ρ to $\widetilde{\rho}$, and that this effect is strongest for cases completed latest.

References

Allen, David (2001). Getting Things Done. Viking.

Angrist D. J., and J. S. Pischke, 2008. Mostly Harmless Econometrics. An Empiricist's Companion. Princeton University Press.

Amemiya, T., 1982. Two Stage Least Absolute Deviations Estimators. Econometrica, 50, 689-711.

Ameriks, J., Caplin, A., and J. Leahy, 2003. Wealth Accumulation And The Propensity To Plan. Quarterly Journal of Economics, 118, 1007-1047.

Aral, S., Brynjolfsson, E., and M., Van Alstyne, 2007. Information, Technology and Information Worker Productivity. NBER WP., 13172.

Bandiera, O., Guiso, L., Prat, A., and R. Sadun, 2008. What CEOs do. Manuscript, London School of Economics.

Victoria Bellotti, Brinda Dalal, Nathaniel Good, Peter Flynn, Daniel G. Bobrow and Nicolas Ducheneaut (2004) What a To-Do: Studies of Task Management Towards the Design of a Personal Task List Manager. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp.735-742, April 24-29, 2004, Vienna, Austria.

Bertrand, M., and A., Schoar, 2003. Managing with Style: The Effect of Managers on Firm Policies. Quarterly Journal of Economics, 118, 1169-1208.

Bloom, N., and J., Van Reenen, 2007. Measuring and Explaining Management Practices Across Firms and Countries. Quarterly Journal of Economics, 122, 1351-1408.

Bloom, N., Propper, C., Seiler, S., and J., Van Reenen, 2009. The Impact of Competition on Management Practices in Public Hospitals. Manuscript, London School of Economics.

Bohn, Roger (2000) Stop Fighting Fires. Harvard Business Review, 78(4) (July-Aug 2000), pp. 83-91 .

Chernozhukov, V., and C., Hansen, 2008. Instrumental Variables Quantile Regression: A Robust Inference Approach, Journal of Econometrics, 142, 379-398.

Chesher, A., 2003. Identification in Nonseparable Models, Econometrica. 71, 5, 1405-1441

Covey, Stephen (1989). The seven habits of highly effective leaders. New York: Simon & Schuster.

Cob, C. W., and P. H., Douglas, 1928. A theory of production, American Economic Review, March 18(2), 139-165.

Coviello, D., Persico, N., and A., Ichino, 2011. Task juggling. Manuscript, New York University.

CSM, 2010: CONSIGLIO SUPERIORE DELLA MAGISTRATURA, Quarta Commissione. Gruppo di lavoro per la individuazione degli standard medi di definizione dei procedimenti. Relazione 2010.

Della Vigna, S., 2009. Psychology and Economics: Evidence from the field, Journal of Economic Literature, 47, 315-372.

Dewatripont, Mathias, Ian Jewitt, and Jean Tirole, 1999. The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies. The Review of Economic Studies, 66(1) pp. 199-217.

Garicano, L., and P., Heaton, 2010. Information Technology, Organization, and Productivity in the Public Sector: Evidence from Police Departments. Journal of Labor Economics, 28, 167-201.

Gibbons, R., and J., Roberts, 2010. Handbook of Organizational Economics. Princeton University Press.

González,Victor M. and Gloria Mark, 2005. Managing Currents of Work: Multi-tasking Among Multiple Collaborations. In H. Gellersen et al. (eds.), *ECSCW 2005: Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work*, 18-22 September 2005, Paris, France, 143–162.

Holmstrom, Bengt and Paul Milgrom, 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, and Organization, Vol. 7.

Ichniowski, C., Shaw, K., and G., Prennushi, 1997. The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. American Economic Review, 87, 291-313.

Jackson, James R., 1963. Jobshop-like Queueing Systems. Management Science, Vol. 10, No. 1 (Oct., 1963), pp. 131-142.

Jorgenson Dale W., 1986. Econometric methods for modeling producer behavior. in "Handbook of Econometrics" Volume III, Edited by Zvi Griliches and M. D. Intriligator, Elsevier Science Publisher.

Lusardi, A., and O., S., Mitchell, 2008. Planning and Financial Literacy. American Economic Review: Papers and Proceedings, 98, 413-417.

Mark, G., D. Gudith, and U. Klocke, 2008, The Cost of Interrupted Work: More Speed and Stress. In *CHI '08: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 107–110, Florence, Italy, April 2008. ACM Press.

Moder,J, C Phillips, E Davis (1983). *Project management with CPM, PERT and project diagramming.* Van Nostrand Reinhold Publishers.

Perlow, Leslie (1999). The time famine: Toward a sociology of work time. Administrative Science Quarterly, 44 (1), pp. 57–81

Repenning, Nelson (2001) Understanding fire fighting in new product development. Journal of Product Innovation Management 1, pp. 85–300.

Stock, J., H., and M., Yogo, 2005. Testing for Weak Instruments in Linear IV Regression. Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, 80-108.

Table 1: The panel structure

Judge identifier	Number of quarters of service per year						Total number of quarters of service	Average number of new cases per quarter
	2000	2001	2002	2003	2004	2005		
1	4	4	4	4	4	0	20	107
3	4	4	1	0	0	0	9	105
5	4	4	4	4	4	4	24	143
6	4	4	4	4	4	0	20	129
7	4	4	4	4	4	2	22	118
8	4	1	4	4	4	0	17	119
9	4	4	1	0	0	0	9	110
10	4	4	4	2	0	0	14	118
11	4	4	4	4	4	4	24	141
12	4	4	4	2	4	4	22	138
13	4	4	4	4	4	2	22	120
14	4	4	4	2	0	0	14	125
15	4	4	4	4	4	0	20	127
18	0	0	0	0	4	4	8	152
19	2	4	4	4	2	4	20	122
20	4	4	4	4	4	4	24	137
21	4	4	4	4	4	4	24	120
22	4	4	4	4	4	4	24	138
24	4	4	4	4	4	4	24	135
29	0	0	0	2	4	4	10	150
30	0	0	0	3	4	4	11	121
Total (average in last col)	70	69	66	63	65	48	381	128

Table 2: Variability of assignments per quarter across judges

Quarter of observation	New cases per judge		Number of judges
	Average	St. Dev.	
2000q1	129	13	18
2000q2	112	11	18
2000q3	82	7	17
2000q4	120	22	17
2001q1	137	20	17
2001q2	134	11	17
2001q3	120	14	17
2001q4	123	21	18
2002q1	134	30	18
2002q2	149	19	16
2002q3	100	11	16
2002q4	144	17	16
2003q1	147	19	16
2003q2	139	21	16
2003q3	108	12	15
2003q4	131	29	16
2004q1	139	17	15
2004q2	151	23	16
2004q3	108	23	17
2004q4	114	31	17
2005q1	123	28	13
2005q2	155	43	13
2005q3	132	18	11
2005q4	161	33	11
Average	128	28	17

Table 3: Tests for the random assignment of cases to judges

Quarter of observation	Type of controversy	Zip code of plaintif's lawyer	Number of involved parties	Number of Judges
2000q1	.089	.052	.003	18
2000q2	.003	.095	.065	18
2000q3	.230	.150	.039	17
2000q4	.045	.015	.000	17
2001q1	.430	.000	.330	17
2001q2	.000	.610	.420	17
2001q3	.760	.670	.660	17
2001q4	.770	.610	.830	18
2002q1	.032	.140	.410	18
2002q2	.130	.570	.270	16
2002q3	.048	.180	.270	16
2002q4	.008	.057	.016	16
2003q1	.720	.410	.410	16
2003q2	.620	.770	.000	16
2003q3	.350	.058	.400	15
2003q4	.120	.098	.033	16
2004q1	.850	.470	.780	15
2004q2	.950	.800	.950	16
2004q3	.190	.100	.040	17
2004q4	.140	.340	.960	17
2005q1	.580	.230	.095	13
2005q2	.004	.810	.450	13
2005q3	.660	.430	.360	11
2005q4	.160	.510	.490	11
N. of rejections	7	2	7	.
Largest balanced panel	.11	.22	.073	.

Note: The top part of this table reports, for each quarter, the p-value of a Chi-square test of indipendence between the identity of judges and three discrete characteristics of cases: type of controversy (14 types); zip code of the platintif's lawyer (55 codes); the number of parties in trial (capped at 10). The central part of the table reports the number of quarters in which independence is rejected at the 5% level. The bottom part of the table reports similar Chi-square tests as in the top part, for all cases assigned in the period spanned by the largest balanced panel of judges identifiable in our sample. As shown in Table 1 this largest panel involves 14 judges observed continuously between year 2000 and year 2002.

Figure 1: Differences of performance between judges with randomly assigned workload

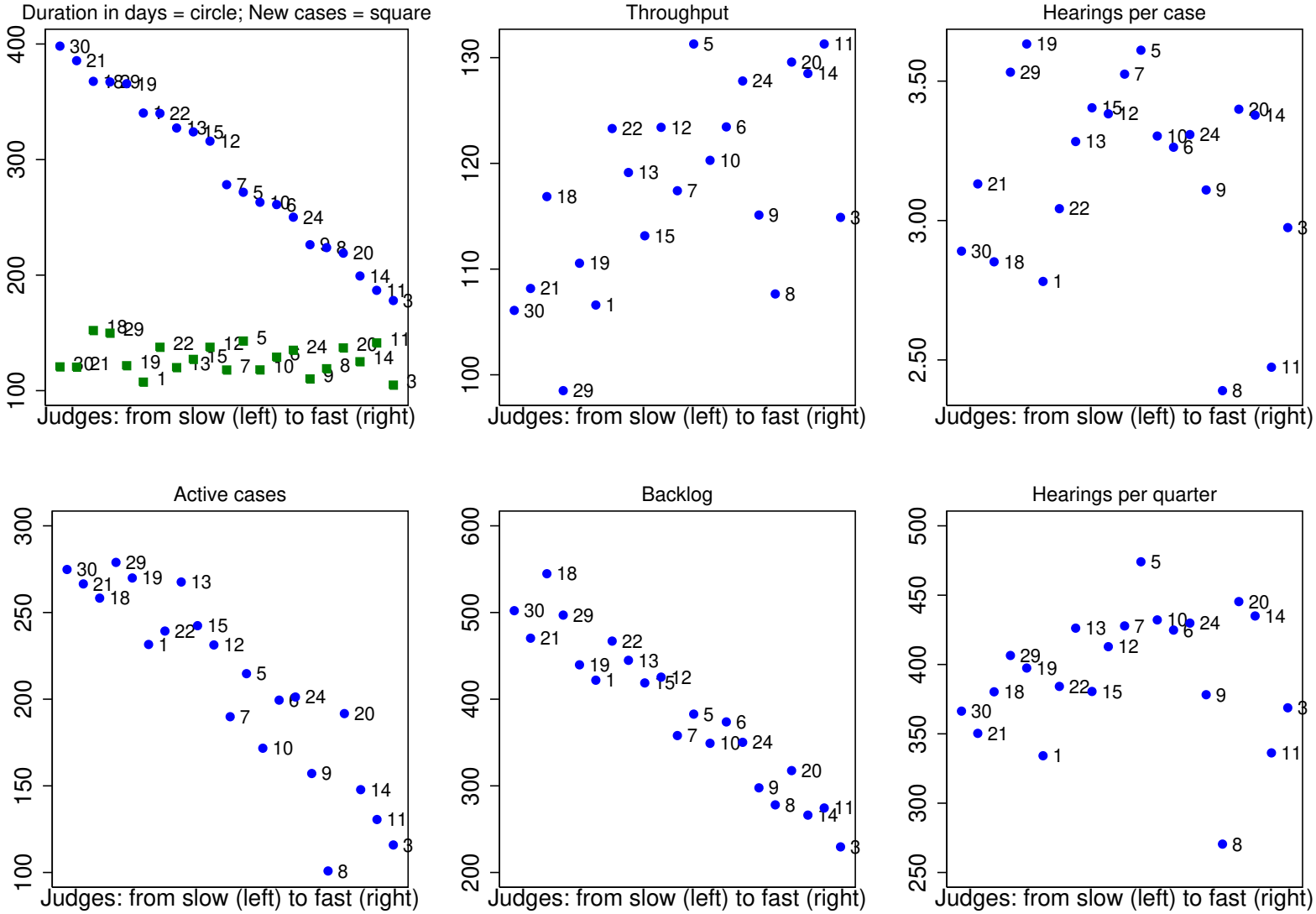

Figure 2: The trade off between quantity and quality in the decision of judges

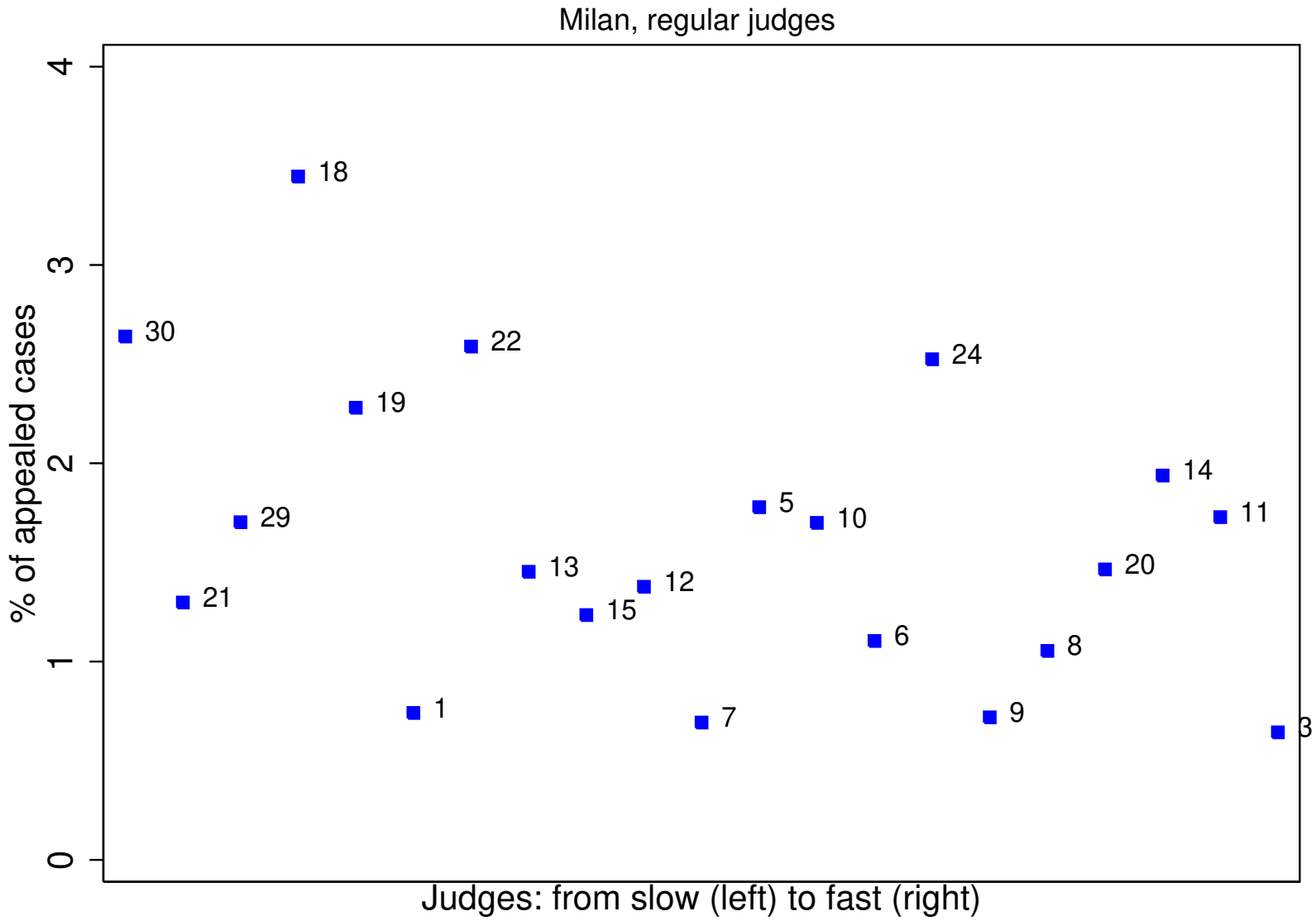

Figure 3: Work flow in a stable rotation

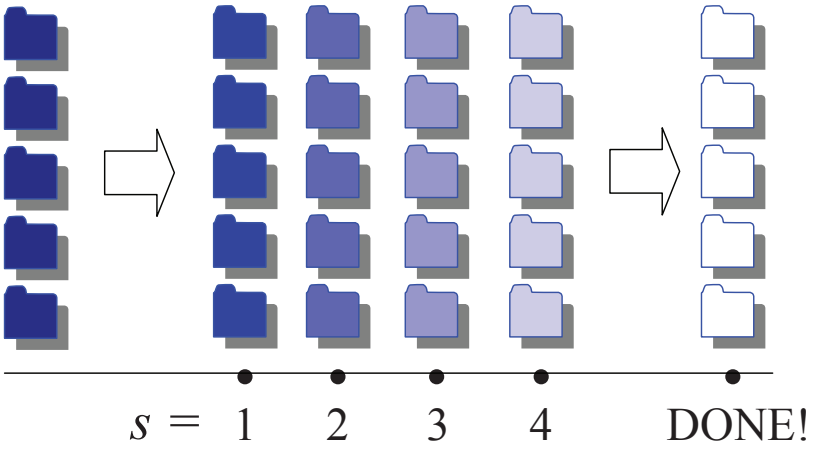

Figure 4: How far are judges from a stable rotation?

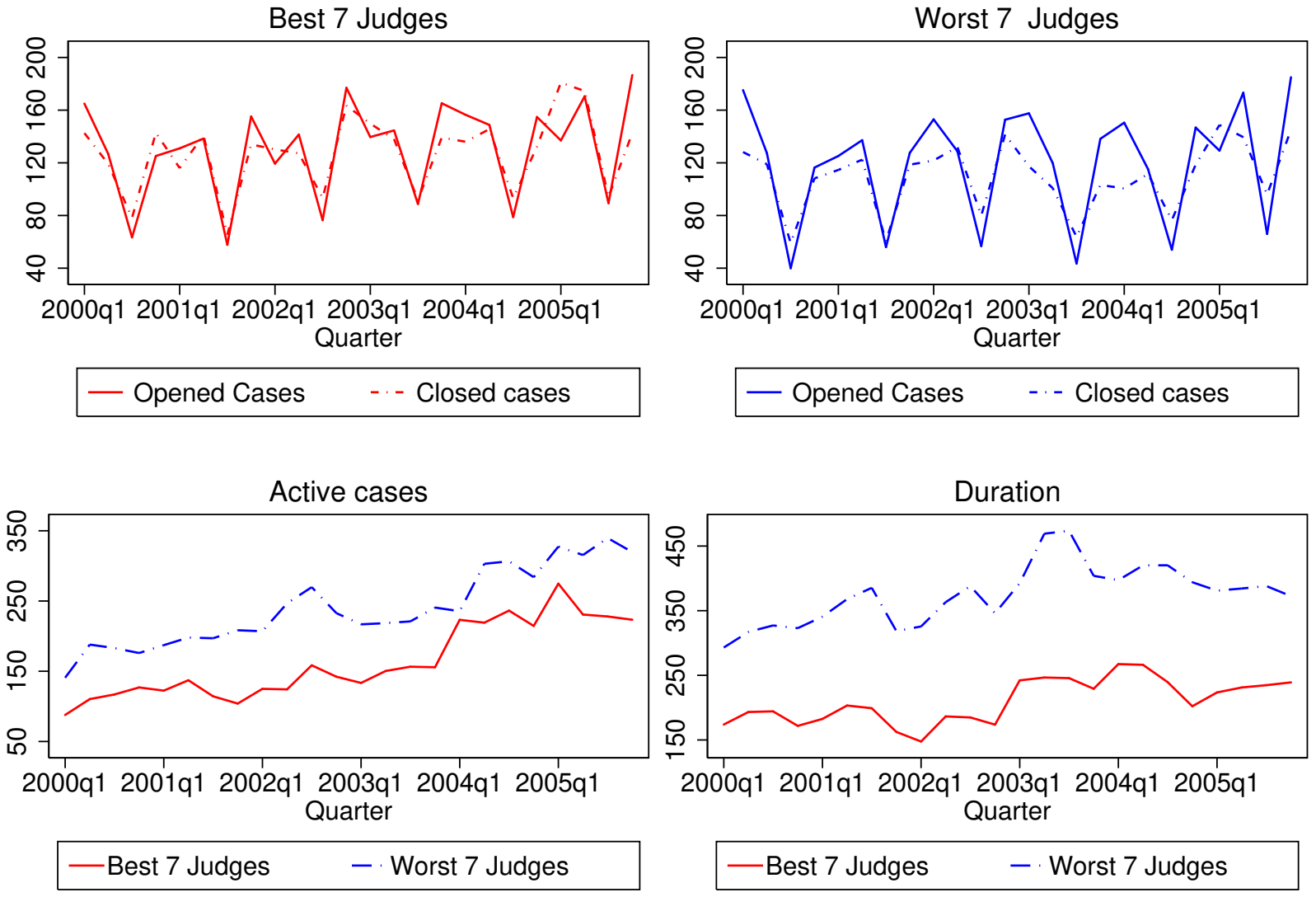

Figure 5: Deviation from a stable rotation

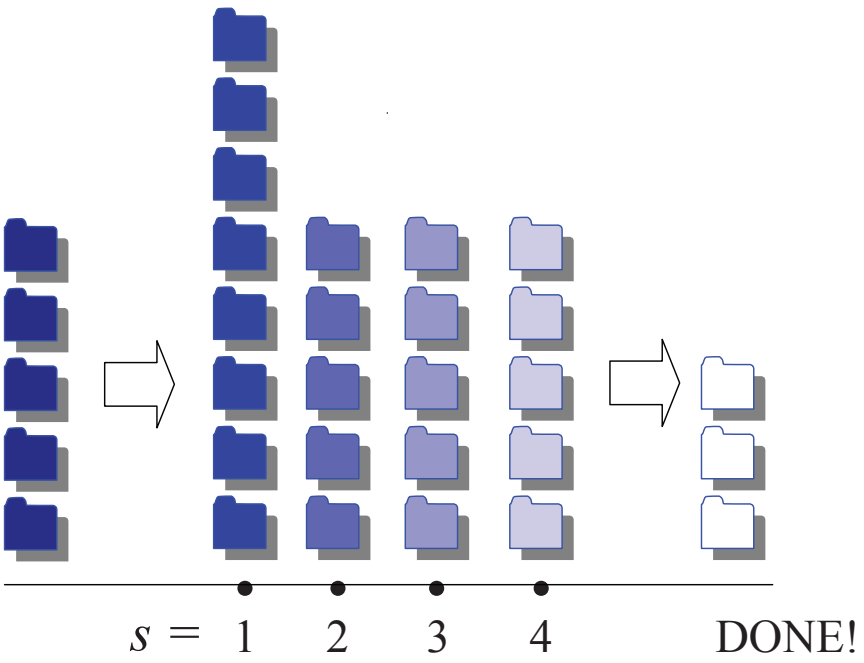

Figure 6: The "60 days" rule and the distribution of first hearings by vintage of cases

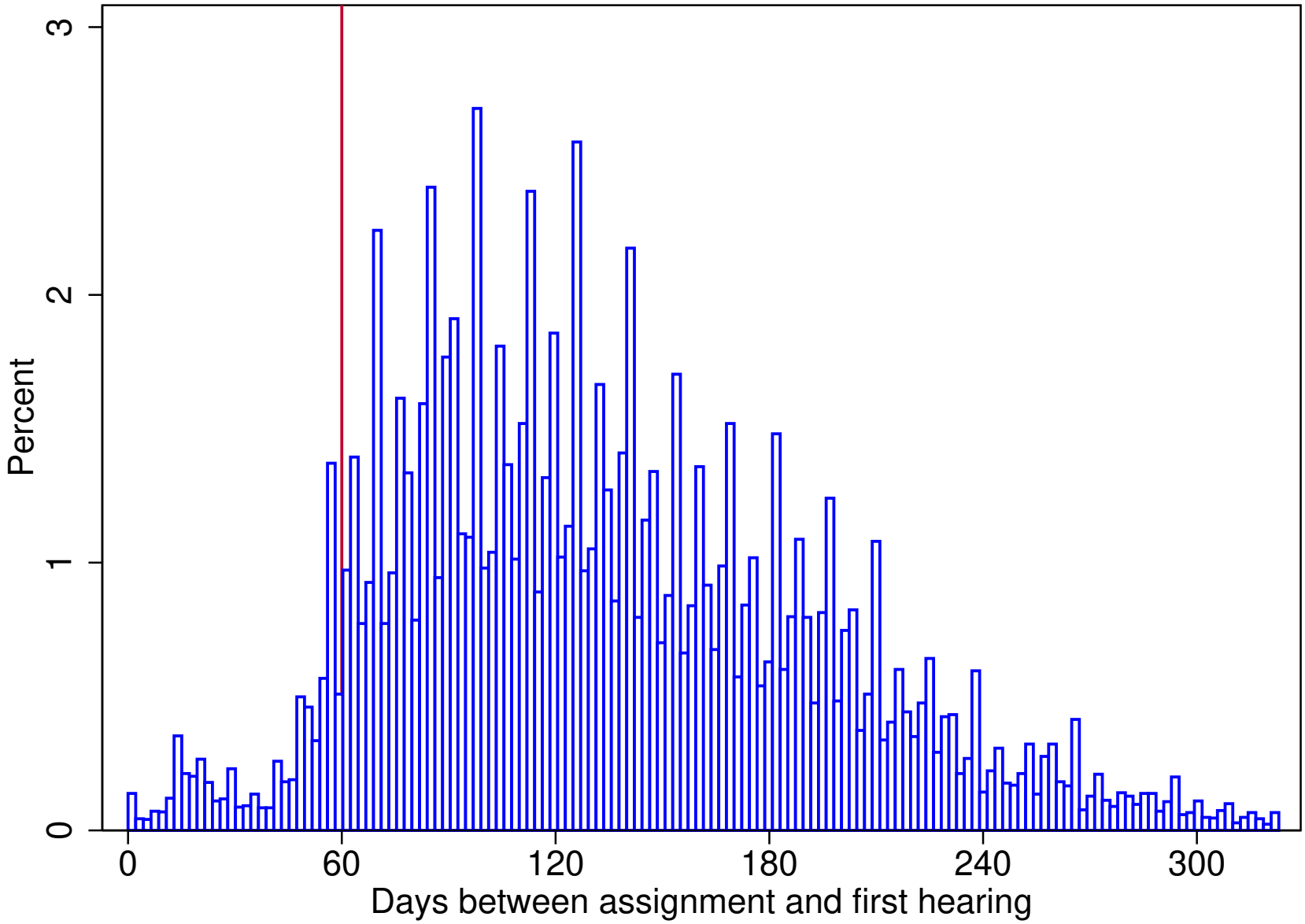

Table 4: Descriptive statistics

	Mean	sd	p25	p50	p75	n
Outcomes:						
Total duration	290	77	229	285	352	381
Total duration of the cases in the highest decile	766	217	618	753	906	381
Inactive duration	127	36	98	122	150	381
Fraction of total duration that is inactive	45	10	38	43	50	381
Closed cases per quarter	119	35	98	122	145	381
Task juggling:						
Active cases at the end of quarter	210	74	154	206	266	381
New opened cases per quarter	127	46	93	137	159	381
Workload, effort, and ability:						
New assigned cases per quarter	128	28	111	132	146	381
Hearings per quarter	396	125	306	425	490	381
Hearings per case	3.2	.57	2.9	3.3	3.6	381
Standardized effort per quarter	128	45	98	131	156	381
Composition of the workload:						
Fraction of new urgent cases assigned per quarter	16	8.1	9	17	22	381
Fraction of new difficult cases assigned per quarter	12	7.1	5.7	12	19	381
Fraction of cases beyond the “60 days limit”	51	6.5	47	51	55	381

Note: All variables are defined per quarter. Standardized effort is defined as the ratio between the Hearings per quarter and the Hearings per case and can be interpreted as the potential number of trials that a judge could complete in a quarter given her average number of hearings per case.

Table 5: The effect of task juggling on average duration

Estimation Method	OLS	IV	OLS	IV
Variables	(1)	(2)	(3)	(4)
$\nu_{i,q}$: New opened cases per quarter	0.39	0.86		
	(0.08)	(0.37)		
$A_{i,q}$: Active cases at the end of quarter			0.29	0.62
			(0.07)	(0.30)
$\alpha_{i,q}$: New assigned cases per quarter	0.31	0.05	0.35	0.18
	(0.09)	(0.19)	(0.09)	(0.15)
$\frac{e}{S}_{i,q}$: Standardized effort per quarter	-0.84	-1.81	-0.67	-1.18
	(0.09)	(0.42)	(0.08)	(0.30)
Implicit trend	4.25	6.67	1.95	1.04
	(0.60)	(1.12)	(0.58)	(1.23)
Cragg-Donald Wald F statistic *(Joint)*		15.19		8.88
Sargan test *(p-value)*		0.93		0.28
Judges fixed effects	Yes	Yes	Yes	Yes
Quarters fixed effects	Yes	Yes	Yes	Yes
Observations	381	381	381	381
Number of Judges	21	21	21	21
R^2	0.54	0.34	0.55	0.43
R^2 including judges' fixed effects	0.85	0.79	0.85	0.81

Note: Robust standard errors are in parentheses (see footnotes 20-24 for a discussion of how they were computed). Standardized effort is defined as the ratio between the hearings per quarter and the hearings per case and can be interpreted as the potential number of trials that a judge could complete in a quarter given his average number of hearings per case. The "Cragg-Donald Wald F statistic *(Joint)*" denotes the minimum eigenvalue of the joint first-stage F-statistic matrix. When denoted with "Yes", regressions include Judges Fixed Effects (21 dummies) and Quarter dummies (2000q1-2005q4).

Table 6: First stage

Endogenous dependent variable Variables	$\frac{e}{S}$ (1)	ν (2)	A (3)
Fraction of new urgent cases assigned per quarter	2.10	1.02	-0.92
	(0.48)	(0.31)	(0.65)
Fraction of new difficult cases assigned per quarter	-0.94	-0.06	-0.18
	(0.56)	(0.40)	(0.61)
Fraction of cases beyond the "60 days limit"	0.37	2.29	2.50
	(0.65)	(0.57)	(0.98)
New assigned cases per quarter	0.02	0.70	0.71
	(0.15)	(0.09)	(0.24)
Cragg-Donald Wald F statistic *(Joint)*		15.19	8.88
Judges fixed effects	Yes	Yes	Yes
Quarters fixed effects	Yes	Yes	Yes
Observations	381	381	381
Number of Judges	21	21	21
R^2	0.77	0.79	0.72
R^2 including judges' fixed effects	0.76	0.79	0.84

Note: Robust standard errors are in parentheses (see footnotes 20-24 for a discussion of how they were computed). Standardized effort is defined as the ratio between "hearings per quarter" and "hearings per case" and can be interpreted as the potential number of trials that a judge could complete in a quarter, given his average number of hearings per case. The "Cragg-Donald Wald F statistic *(Joint)*" denotes the minimum eigenvalue of the joint first-stage F-statistic matrix. The F statistic reported in column 2 is for $\frac{e}{S}$ and A, while the one reported in column 3 is for $\frac{e}{S}$ and ν. When denoted with "Yes", regressions include Judges fixed effects (21 dummies) and Quarter dummies (2000q1-2005q4).

Table 7: Testing for the Presence of Switching Costs or Benefits

Panel A: Standardized Effective Effort ($\frac{\widetilde{e}}{S} = \frac{\frac{e}{(1+\phi P)}}{S}$)

	OLS	IV	OLS	IV
New opened cases per quarter	0.50	0.19		
	(0.05)	(0.22)		
Active cases at the end of quarter			0.04	0.17
			(0.05)	(0.21)

Panel B: Percentile of the Duration distribution of trials

90^{th} Percentile	QREG	IVQREG	QREG	IVQREG
New opened cases per quarter	0.04	0.19		
	(0.15)	(1.08)		
Active cases at the end of quarter			0.05	1.09
			(0.15)	(1.32)
95^{th} Percentile	QREG	IVQREG	QREG	IVQREG
New opened cases per quarter	-0.15	-0.82		
	(0.16)	(1.59)		
Active cases at the end of quarter			0.05	0.12
			(0.15)	(1.39)
Number of Judges	21	21	21	21
Observations	381	381	381	381

Note: Panel A reports the effects of the two indicators of task juggling ($P_{i,q} = \nu_{i,q}$ and $P_{i,q} = A_{i,q}$) on standardized effective effort, estimated with equation (9) in the text. Panel B reports instead the effects of the same variables on the 90th and 95th percentiles of duration distribution, estimated with equation (10) in the text. These quantile and quantile instrumental variables estimates have been obtained with the Least Absolute Deviations estimator (QREG) and the Chernuzkov and Hansen (2008) Instrumental Variables Estimator (IVQREG). We are grateful to Do Wan Kwak who shared with us his Stata code that implements the Chernuzkov and Hansen (2008) estimator. In Panel A, standard errors are computed with the robust formula (see footnote 20-24 for a justification of this choice). In Panel B standard errors are bootstrapped. All the regressions include judges fixed effects (21 dummies) and quarter dummies (2000q1-2005q4) and control for the quarterly workload.