Don't spread yourself too thin
The impact of task juggling on workers' productivity*

Decio Coviello
EUROPEAN UNIVERSITY INSTITUTE

Andrea Ichino
UNIVERSITY OF BOLOGNA

Nicola Persico
NEWYORK UNIVERSITY

April 6, 2010

Abstract

We show that task juggling, i.e., the spreading of effort across too many active projects, decreases the performance of workers, raising the chances of low throughput, long duration of projects and exploding backlogs. Individual productivity cannot be explained only in terms of effort, ability and experience: work scheduling is a crucial "input" that cannot be omitted from the production function of individual workers. We provide a simple theoretical model to study the effects of increased task juggling on the duration of projects. Using a sample of Italian judges we show that those who are induced for exogenous reasons to work in a more parallel fashion on many trials at the same time take longer to complete similar portfolios of cases. The exogenous variation that identifies this causal effect is constructed exploiting the lottery that assigns cases to judges together with the prescription requiring judges to hold the first hearing of a case no later than 60 days from filing.

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

*We would like to thank seminar participants at the NBER Summer Institute 2009, the WPEG 2009 conference, Bocconi University, European University Institute, University of Bologna, Timbergen Institute, Utrecht University. We are also grateful to the Labor Court of Milan for making the data available to us.
Email: andrea.ichino@unibo.it; decio.coviello@gmail.com; nicola@nicolapersico.com

1 Introduction

Consider a worker that receives the assignment of two independent jobs, A and B, each requiring 10 days of full dedication to be completed. If she juggles both jobs, for example dealing with A on odd days and with B on even days, the average duration of the two tasks is equal to 19.5 days. If instead she focuses on each job in turn, completing A alone in the initial 10 days and shifting to B only later in the next 10 days, the average duration of both jobs from the time of assignment is 15 days. Note that under the second work schedule job B does *not* take longer to complete, while A is completed much faster; in other words, avoiding task juggling results in a Pareto-improvement across task durations. This simple example suggests that, conditional on effort, ability, and size of assigned workload, workers who juggle too many tasks at the same time may take longer to complete them.

In this paper we study theoretically and empirically the hypothesis that task juggling decreases the performance of workers raising the chances that they enter in a situation of congestion and overwhelm characterised by low throughput, long duration of tasks completion and exploding backlogs. Our results suggest that individual productivity cannot be explained only in terms of effort, ability and experience. Individual work scheduling (how much juggling is done) is a crucial input that cannot be omitted from the production function of individual workers.

Using a sample of Italian judges who receive a randomly assigned workload (Section 2), we show that the heterogeneity of their performance is considerable and cannot be fully explained in terms of measures of experience, ability and effort, even if these measures are very precise and error-free in our data (Section 3). Descriptive evidence suggests that judges who keep only few trials active, trying to close them as quickly as possible before starting new ones, dispose more rapidly of a larger number of cases per unit of time. In this way, their backlog remains low even though they receive the same workload of other judges who juggle more trials at any given time.

To rationalize this evidence, we propose a theoretical model that explains how parallel and sequential tasks scheduling affect performance in terms of duration, throughput and backlog (Section 4). The model, which builds on Persico et al. (2010), also suggests an explanation of why judges choose to juggle different number of cases. This explanation is based on the idea that parties in trial lobby to have their case dealt sooner by the judge. This lobbying behaviour is privately optimal for the lobbying parties, but socially inefficient because judges who cannot resist this multilateral pressure increase the number of cases they juggle, slowing down the completion of all assigned cases. Thus, the heterogeneity in the

performance of judges, for given effort and ability, ultimately depends on the heterogeneity in their capacity to resist the lobbying of parties in trials. Although described for the specific setting of the judges considered in the empirical analysis, the suggested mechanism that generates task juggling applies more generally to environments in which a worker interacts with different "customers" or "principals" waiting for him to complete a set of assigned jobs.

We then use regression analysis on the already mentioned panel of Italian judges, observed for six years, to show that the choice of work scheduling has quantitatively important effects on performance, compared to variation in experience, ability and effort (Section 5). In order to identify the causal effect of parallel tasks scheduling (i.e. task juggling) we construct time-varying instruments 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 number of assigned cases generate pressure for more parallel working around and after 60 days from filing. Results strongly support the hypothesis that judges induced for exogenous reasons to work in a more parallel fashion take more time to complete similar portfolios of cases.

The final Section 6 concludes discussing results and their implications.

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 have information also 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 performance starting from the fifth of their quarters of service, in order to give them time to settle. 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 performance. Thus all the judges that we analyze have at least one year of tenure and we know their backlog of unsettled 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.

As in other countries, also in Italy, the Law (Art. 25 of the Constitution) requires that judges receive a randomly assigned portfolio of new controversies in order 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 this prescription is implemented in the following way. Every morning judges in service are ordered alphabetically, starting from a randomly extracted letter of the alphabet and the cases filed during the day are assigned in sequence to all judges in service. As a result, given the large number of new cases filed per quarter, the portfolios of controversies is qualitatively similar for all the judges that we observe within each quarter. Remaining differences across judges are due to random variability of assignments and are independent of the identity of judges. For example, if in a given day the letter extracted is B and 5 cases are filed, only judges with a name starting from B to F will receive an assignment on that day.

At the end of this selection process we are left with the subset of judges described in Tables 1 and 2. Of the original 31 judges, 21 have a quantitatively and qualitatively comparable workload (up to random differences) *within each quarter* and are therefore retained for the analysis. Table 1 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 similarly small in the other quarters. Table 2 reports the number of quarters in which each judge is observed over each year and in total. The panel is unbalanced, with 6 judges observed for all the 24 quarters, while the others are observed for fewer quarters with a minimum of 8 quarters. The last column of Table 2 reports the number of cases assigned to each judge per quarter on average.

For the purpose of this study it is important to realize that the workload of a judge may change across quarters because of the temporal variation in the number of filed cases or in the number of judges in service, and therefore for reasons that are independent of the identity of judges. Moreover, within each quarter, the qualitative and quantitative workload may differ across judges because of the variability randomly generated by the assignment procedure described above and thus once again independently of the identity of judges.

As we will see, for the purpose of identification 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 than in other datasets that are typically richer of demographic characteristics but do not offer 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 performance, 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 the filing date 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.[1] 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. 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

[1] See Section 2.

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 very similar workload in terms of quantity and quality.

3.2 Throughput and backlog

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. 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 best performer 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).

Our analysis suggests that the capacity of a judge to work on only few cases at the same time, independently of how many have been assigned to her, is likely to improve significantly

her capacity to dispose quickly of a large number of cases and to avoid an uncontrolled explosion of the backlog. In other words, sequential working, as opposed to parallel working, helps to avoid a situation of congestion and overwhelm.

3.3 Complication of cases, ability and effort of judges

Although suggestive, our hypothesis concerning the role of parallel working on the performance of judges must be confronted with other more obvious potentially relevant determinants of this performance. In this section we focus on proxies of 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 day 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 and throughput. Interestingly, the slowest two Judges 30 and 21 schedule the same number of hearings per day than the fastest two judges 3 and 11. If anything, it seems that the capacity to "keep fewer pots on the fire" allows the

faster judges to economize on effort, but all in all the correlation between this measure of effort and duration is practically null (-0.06)

In other words, despite the fact that the performance of judges, in terms of duration and throughput, is very heterogeneous, no clear strong link emerges between this performance and good proxies of ability and effort like the number of hearings needed to handle cases of similar complication and the number of hearings per unit of time.

To summarize, the descriptive evidence presented in this section suggests that parallel working, 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. The judges who work on few cases at the same time and try to close them quickly before opening new ones, succeed in closing a larger number of cases per quarter and in less time from the assignment. In this way they can keep low the backlog at the beginning of each quarter, even if they receive the same number of cases per quarter as their slower court-mates. These latter, who tend to work in parallel on many cases, appear to be overwhelmed and their performance suffers. 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). The judges who perform better in terms of throughput and duration of cases seem to be also the ones who take decisions of better quality.

4 Theory

4.1 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 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.[2]

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.

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

[2]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.

Definition 1. *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 2. *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 (see the Appendix 7.1) 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. ***(Sequential schedule)*** *Let there be three cases each requiring two steps, so that $C = 3$ and $S = 2$. The sequential work schedule is*

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

(Full rotation schedule) *Let there be three cases each requiring two steps, so that $C = 3$ and $S = 2$. The full rotation work schedule is*

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

(Rotation on the open) *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.

4.2 Effect of Parallel Work on Durations

In this section we show that a judge who works more "in parallel" takes more time to complete *all* his cases. 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 3. *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 4. *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.

We now introduce the main theoretical result of the paper, showing that the judge who opens many cases early completes them all late.

Proposition 1. *Suppose C is finite. Suppose that, within the family of rotations on the open, we take a specific schedule and change it by anticipating the opening date of case $\widehat{c}$ and all following cases. Then every case is completed no earlier, and some are completed strictly later.*

Proof. See the Appendix 7.1. □

This proposition is the main theoretical insight in this paper. It says that anticipating the opening of a case 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.

We now extend the logic of this proposition to the situation in which cases are heterogeneous in their length S_c. Rather than developing a full-blown theoretical model of heterogeneous cases, we limit ourselves to showing that the main result of this section, namely, that a specific sequential schedule is faster than a full rotation, is maintained. To be precise, however, we now need to realize that there are many sequential schedules, each characterized by the prioritization of cases with different S_c. The duration-minimizing schedule is the one in which cases are worked on one at a time (sequential), and the priority is such that if $S_c < S_{c'}$ then case c must be started (and completed) before case c' is touched. All other schedules, including the full rotation one, give a larger total duration. The logic is most easily seen via

an example. Suppose we have two cases c and c', both assigned at time zero. Let $S_c = 5$ and $S_{c'} = 10$, so case c takes fewer steps to complete. If we schedule the cases sequentially starting from c (the shortest one), the sum of total durations is $5 + 15 = 20$. If we do them in parallel (full rotation starting with step c_1) the sum of total durations is $9 + 15 = 24$. This example shows that *a sequential schedule which prioritizes short cases* is faster than a parallel schedule. It is easy to convince oneself that this principle holds in general, no matter what the number of cases and their individual complexity S are. Moreover, the same principle applies if cases are not all assigned at zero, but rather some are assigned while the judge is in mid-process. In this case the duration-minimizing strategy is the following. At each point in time the judge should evaluate the length of each case in front of her and work only on the case with the fewer steps to completion. This is again a sequential work schedule, but one that allows for newly assigned cases to "cut in line" and be worked on if they have the fewest number of steps to completion. If a case "cuts in line" then the case previously being worked on should be kept on hold until it again becomes the case with the fewer number of steps to completion.

Bottom line: also in the presence of heterogeneity and different assignment dates, if the goal is just to minimize the duration of cases the optimal work schedule is a sequential schedule with only one open case at any moment. The single open case on which the judge should work would have to be the one closer to completion within the existing backlog. If other objectives suggest different orderings of the sequence of cases or a rotation of effort on more than one case at the same time, it should be clear that reaching these objectives would imply the cost of increasing average duration.

4.3 Towards a Theoretical Framework for the Empirical Analysis of Congestion

In this section we want to obtain an expression for the duration of a case, as a function of several inputs: the effort, the complexity of cases, the arrival rate of new cases and the degree of parallelism, or equivalently the number of active cases at any point in time, which measures the congestion the judge operates under.

The effect of the first two variables, effort and complexity, can be appreciated even in the most stark model in which there is only one case, $C = 1$. This case 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 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:[3]

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

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

When we have more than one case, a third factor beyond e and S affects the duration of cases, namely, how many cases the judge keeps open at any point in time, which is a measure of congestion. The easiest way to generalize equation (1) so as to account for the effect of congestion is to study a system that evolves through time, but that does so in a very stable way. To this end we now introduce the simplest possible evolution of the system over time.

Definition 5. *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 of cases S, the effort e, and the number of new cases opened, are all constant in each quarter;

(c) the work schedule is a rotation on the open;

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

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 open cases that have had one hearing, 5 open 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 open case; this is because the judge follows a rotation on the open. 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

[3] 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.

Appendix 7.2 we prove that the input rate and the output rate must be exactly equal to e/S for there to be a stable rotation.

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 7.2 we fully analyze how a stable rotation operates and obtain the following expression for the duration of cases.

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

This expression solves for the duration D_q of cases assigned in quarter q in terms of the 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 2. *In 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.*

Proposition (2) provides a theory-based starting point for implementing an econometric analysis of the contributing factors to durations. However, Proposition (2) is limited in an important way: it describes a kind of "stable state" in which cases are opened and closed at the same rate. But are the judges considered in this study effectively working under such a stable rotation? We address this question in the next Section 4.4, where we show that although they are close to a stable rotation, their opening rate of cases is not constant over time and is often larger than their closing rate. 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 Section 4.5, to the more interesting and realistic case in which congestion can change.

4.4 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:[4]

$$\underset{}{\nu} = \underset{(5.42)}{7.96} + \underset{(0.04)}{1.00}\,\omega \tag{3}$$

where standard error are reported in parentheses under the coefficients. According to these estimates these judges work on a schedule that is very close to stable rotation but does not coincides exactly with it. The slope equal to 1 indicates that judges open one new case for each case that they close. But the positive intercept (even if statistically not different from 0) suggests that on average they (or possibly some of them) also open approximately 8 new cases in every quarter on top of those that they close. As a result the number of active cases on their desk steadily increase 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 times 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

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

effects on the number of active cases and on the duration of assigned cases. We therefore have to incorporate in our theory of the production function of judges also the effects of deviations from a stable rotation and more generally of changes in the number of opened cases. To this end is devoted the next section.

4.5 What happens outside of a stable rotation

To capture the notion of an increase in congestion we need to think of a (temporary) increase in the number of cases newly opened in quarter q, like, for example, the one described in Figure 5. If we denote this number by ν_q, then in a stable rotation we have $\nu_1 = \nu_2 = ... = \nu$. Suppose now we increase ν for a specific quarter, say we double ν in quarter 3. In other words we start from the steady-state pattern $\nu, \nu, \nu, ...$ and we want to check the effect on durations of going to the pattern $\nu, \nu, 2\nu, \nu, ...$ What is the effect of increasing congestion? By Proposition 1, the effect will be to increase durations.[5] Therefore, we have the following proposition, which is really a corollary of Proposition 1.

Proposition 3. *Starting from any work schedule (including a stable rotation), increasing ν_q, the number of cases newly opened in quarter q, while keeping constant standardized effort $\left(\frac{e}{S}\right)_q$, increases the total duration of cases assigned at q.*

What mechanism can generate variation in ν_q? More generally, what do we think determines the judge's choice of how many new cases to open? The question is particularly relevant if we conclude, as we will, that the judges' output rate suffers because of an excessive number of new cases opened in each quarter. In a related paper (Persico et al. 2009), we propose and analyze theoretically a possible mechanisms. The idea is that judges open too many cases because they are under pressure by the parties to do so. In a nutshell, the model is the following. Every day, a judge holds S hearings. The parties of every case assigned to the judge would like their case to have as many hearings as possible on that day. The parties can pay a lobbying cost and pressure the judge to hold as many hearings as possible on their case in that day. Judges are unable to fully resist this pressure and so anyone who wants to bear the cost of lobbying the judge can ensure that, on any given day, the judge devotes an equal proportion of his time to their case. The pressure from multiple parties forces the judge to split his effort across several lobbied cases, which may lead to holding one

[5] Of course, given the restriction to finite C, it becomes necessary for us to have zero cases opened in the last quarter, which would not necessarily be the case in a dynamic model where cases are infinite. But this is not really a limitation, since we prove that the pattern $\nu, \nu, 2\nu, \nu, 0$ produces longer durations than the pattern ν, ν, ν, ν, ν. A fortiori, the pattern $\nu, \nu, 2\nu, \nu, \nu$ (which we do not study) would produce even longer durations, and thus strengthen our conclusions.

hearing on each of S different cases. In other words, the judge is endogenously "forced into" a work schedule that resembles a rotation on the open.[6] Moreover, because the parties with cases assigned but not yet opened can lobby as well, this mechanism results in an excessive amount of cases opened in each quarter. The degree to which judges are subject to this inefficiency depends, among other things, on the judge's power to resist the pressures at any particular time. Formally, this power can be modeled by allowing the lobbying costs to vary across judges and across quarters. Such a variation is a plausible source of variation in ν_q across judges and quarters.

We now have all the elements to specify a theory-based econometric model of the duration of cases, with the goal of estimating the causal effect of an increase in the degree of parallelism of a judge. This is done in the next section.

5 Econometric evidence on the effect of parallel working on trials' duration

5.1 Specification

According to standard theories of the production function of judges, that consider as irrelevant 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. The theory presented in Section 4 suggests, instead, that measures of the extent to which a judge work in parallel must be included in the specification. The simplest way to introduce such measures is offered by equation 2, which is derived under the assumption that the judge works according to a "stable rotation". A linear approximation of equation 2 is

$$D_{i,q} = \gamma_0 + \gamma_1 \alpha_{i,q} + \gamma_2 \left(\frac{e}{S}\right)_{i,q} + \gamma_4 q + \gamma_5 A_{i,0} + u_{i,q} \tag{4}$$

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, as explained in Section 4.3.

[6] Note that, if the judge was not pressurized, he might well be able to focus all S daily hearing on one case, i.e., operate on a sequential work schedule, thus maximizing his output.

However, this specification is still unsatisfactory for two main reasons. First, for some judges we do not observe the initial condition $A_{i,0}$ and thus even if judges were working on a stable rotation we could not estimate the consequences of a higher degree of parallel working inasmuch as this is determined by the initial condition $A_{i,0}$. Second, and perhaps more importantly, we know from Section 4.4 that judges do not operate according to a stable rotation and the number of active cases is not constant over time at the initial value $A_{i,0}$. Outside of a stable rotation, Proposition 3 holds and therefore the correct specification must include a variable $P_{i,q}$ that measures how the degree of parallelism changes with respect to the initial condition. We measure the degree of parallelism in two alternative but related ways: with the variable $A_{i,q}$ which measures the number of active cases on the table of judge i at the end quarter q or with the variable $\nu_{i,q}$ which measures the number of new opened cases by judge i in quarter q.

As a result of these considerations the correct specification that we want 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 (2), but not exactly since Proposition (2) 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 $\beta_1 > 0$. 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.

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. Thus, even controlling for quarters and judges fixed effects, the data display judge specific variability over time of the workload $\alpha_{i,q}$.

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 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 $S_{i,q} > S_{i,p}$ and, with $q < p$, 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.

The main focus of our analysis is on the parameter β_3 which measures the effect of parallelism on duration. Proposition 3 states without ambiguity that this coefficient should be estimated to be positive independently of whether the degree of parallelism is measured by $P_{i,q} = A_{i,q}$ or $P_{i,q} = \nu_{i,q}$.

Finally, Proposition (2) 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.

5.2 Identification

While $\alpha_{i,q}$ is randomly assigned, 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 degree of parallelism $P_{i,q}$, however measured. This because the error term includes lagged and forward values of standardized effort as well as the unobservable parameter that describes the capacity of judges to resist the lobbying of parties in trials who want to anticipate the first hearing of their case. As discussed in Section 4.5, this unobservable parameter ultimately determines the degree of parallelism chosen by a

judge and there is no reason to expect that it should be time invariant, given that it could change with the difficulty of assigned cases or their number.

Therefore to estimate consistently the causal effects of standardized effort and parallelism on trials duration with equation 5, we need some exogenous source of variations of these two variables. As far as standardized effort is concerned, this exogenous source of variation is offered by the alphabetical system 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 number of "urgent" cases and the number of "difficult" cases that judges receive in each quarter.[7]

Note that these instruments, which capture the complexity of assigned controversies conditionally on the size of the workload, can affect the duration of cases only 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 feature can affect duration only if the judge changes the number of hearings held in the 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 degree of parallelism $P_{i,q}$, whether measured with the number of active cases $A_{i,q}$ or with the number of new opened cases $\nu_{i,q}$, can instead be constructed exploiting a prescription that constraints the freedom of judges to decide when to hold the first hearing of a case. Judges are in fact invited to hold the first hearing within 60 days from the filing of a case. There is no penalty for a delay but if 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 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. We therefore construct an instrument for the number of new opened cases defined as the number of cases assigned to the judge in the previous quarter. Note that, as any assignment-to-treatment mechanism, this suffers the possibility of non-compliance. Not all judges feel the pressure of this prescription, but some do feel it, as we will see, and tend to open more or less new cases depending on how many, within the assigned load, have passed the 60 days threshold. Moreover, note that the instrument is randomly assigned

[7]The classification of a case as "urgent" or "difficult" was implemented using an independent survey of judges which were asked to rank the typology of possible cases according to urgency and complication.

because it depends only on the total number of filed cases in the previous quarter and on the specific number of cases received exogenously by each judge as a result of the alphabetically daily assignment system. The instruments displays judge specific variability over time and is therefore compatible with the inclusion of judges and quarter fixed effects.

Thus, 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 assigned cases. Interestingly, when the results of this research were made public in Italy, some judges who were put under investigation because too many of their first hearings took place 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.

5.3 Estimates

Table 3 gives the descriptive statistics for the variables used in the econometric analysis, while results of the estimation of equation 5 are presented in Table 4. In the first column the degree of parallelism $P_{i,q}$ is measured with the number of active cases $A_{i,q}$ on the desk of each judge at the end of a quarter. All estimates are statistically significant and the signs correspond to the prediction of the theory. In particular, a larger number of active cases increase the average duration of cases assigned during the 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. Column 2 reports Instrumental Variable (IV) estimates obtained using the instruments described above in Section 5.2. The effects of the confounded variables $A_{i,q}$ and $\left(\frac{e}{S}\right)_{i,q}$ are now larger and still statistically significant. At the mean of the distribution of active cases (210)[8], a decrease of task juggling consisting of ten fewer active cases (approximately a 5% decrease) reduces by 6.6 days the average duration of all cases assigned in the same quarter, which, given a mean duration of 290 days, is a 2.3% improvement. 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 duration of trials. Given an estimate of -1.29 for the coefficient of $\left(\frac{e}{S}\right)_{i,q}$, 5.1 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 6.6 days. In other words, at

[8]See the descriptive statistics in Table 3

the mean, a 5% decrease of task juggling has the same effect of a 4% increase of effort. If the average number of hearings per case is $S = 3.2$, this means 16.3 more hearings per quarters.

In the third column of Table 4 we report estimates that measure the degree of task juggling with the number of new opened cases $\nu_{i,q}$, i.e. the number of assigned cases for which the judge holds the first hearing. Also in this cases all the estimates are statistically significant and the signs correspond to the prediction of the theory. Using the corresponding IV estimates of the fourth column to compare the size of the effects, ten fewer newly opened cases in a quarter (an 8% decrease of this indicator of task juggling, at the mean of 127 new opened cases per quarter), reduces the duration of assigned cases by 7.9 days (a 3% improvement). To achieve the same effect with more standardized effort per quarter the representative judge would have to increase it by 4.6 units. So, in this case, an 8% decrease of task juggling has the same effect as a 4% increase in standardized effort. If the average number of hearings per case is $S = 3.2$, this means 14.7 more hearings per quarters.

Table 5 reports results from first stage estimations, which show that the instruments we use are sufficiently strong to justify our interpretation of the IV estimates of equation 5. In columns 2 and 3 of table 5 we report the minimum eigenvalue of each of the joint first stage F-statistic (matrix). Both statistics are well above 8.18. This is the critical value computed in Stock and Yogo (2005), table 5.2, for the corresponding linear IV case.

We conclude that the evidence based on the judges considered by this study confirms the prediction of our theory. Judges who are induced 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 of fewer hearings to close a case.

6 Conclusions

We presented theoretical reasons and empirical evidence in favor of the hypothesis that individual work scheduling has significant effects on the speed at which workers can complete assigned jobs. Specifically, we claimed that, for given size of assigned workload, workers who juggle too many tasks are necessarily slower than workers who concentrate sequentially on few tasks at the same time. Relative to our companion paper (Persico et al. 2010), that contains a fully fledged model of work scheduling and performance, the theoretical contribution in this paper is to show that, ceteris paribus, a non-permanent increase in new cases opened in one period increases the duration of the cases that are yet to be completed, regardless

of whether the worker is in a constant growth path. The intuition is that by adding one task to those which the worker is already juggling, she pulls resources away from her other active tasks which are closer to being completed. Moreover, the newly opened task does not benefit from being opened earlier, in the sense that it will still have to wait before all other tasks are completed.

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. In order to identify this causal effect 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.

Our results fit broadly within recent literature suggesting that, in different areas of human behaviour, individual modes of activity scheduling correlate with performance for given effort.[9] But, 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 fairly tightly the causal effect of work practices on performance.

We view the analysis in this paper and its companion (Persico et al. 2010) as a first step into the theoretical and empirical analysis of work scheduling. Although the intuition for the inefficiency of task juggling is strong, measuring the quantitative effects of task juggling is far from straightforward. There are several reasons for this. First, since we are dealing with a dynamic production function, the productivity at each point in time depends on inputs in past periods, which is a source of complexity. Second, work schedules come in an almost infinite range of variations, in principle equal to all the ways in which S steps of each of N tasks can be ordered (a very large cardinality indeed!). Our theoretical approach deals with this complexity by parameterizing work schedules according to a simple measure: how many new tasks are opened in each quarter (with few tasks corresponding to little juggling). Establishing the empirical relevance of this simplification is a large part of the methodological contribution of this paper.

[9] See, for example, Ichniowski et al.(1997) for workers in steel plants, 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. See also Gibbons and Robert (2010) for a recent more general survey.

7 Appendix

7.1 Omitted proofs

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 1 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 1 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 1.

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 2). The other is $c_{s+1} \prec c_{s+2} \prec c'_2 \prec c'_3$, which violates Definition 2 (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.

Proof of Proposition 1.

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. □

7.2 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 steady state they will all be constant over time.[10]

[10]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

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.[11]

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

$$A_0 + \alpha q - \nu q$$

unfinished cases in front of it.[12] 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).

load of the outgoing judge which he replaces.

[11] 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.

[12] 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.

References

Ameriks, J., A., Caplin, and J. Leahy, 2003. Wealth Accumulation And The Propensity To Plan. The Quarterly Journal of Economics, 118, 3, 1007-1047.

Anderson, J., M., Kling J., R., and K., Stith, 1999. Measuring Interjudge Sentencing Disparity: Before and after the Federal Sentencing Guidelines, Journal of Law and Economics, 42, pg.271-307

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

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

Bertrand, M., and A., Schoar, 2003. Managing with Style: The Effect of Managers on Firm Policies. The Quarterly Journal of Economics, 118, 4, 1169-1208.

Bloom, N., and J., Van Reenen, 2007. Measuring and Explaining Management Practives Across Firms and Countries. The Quarterly Journal of Economics, 122, 4, 1351-1408.

Bloom, N., Propper, S., Seiler, and J., Van Reenen, 2009. The Impact of Competition on Management Practices in Public Hospitals. Mimeo

Persico, N., Coviello, D. and Ichino, A., 2010. Task juggling. Manuscript, New York University.

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

Ichniowski, C., K., Shaw, and G., Prennushi, 1997. The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. The American Economic Review, 87, 3, 291-313.

Lusardi, A., and O., S., Mitchell, 2008. Planning and Financial Literacy. American Economic Review: Papers and Proceedings, 98:2, 413417.

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, 80108.

Table 1: 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 2: 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

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

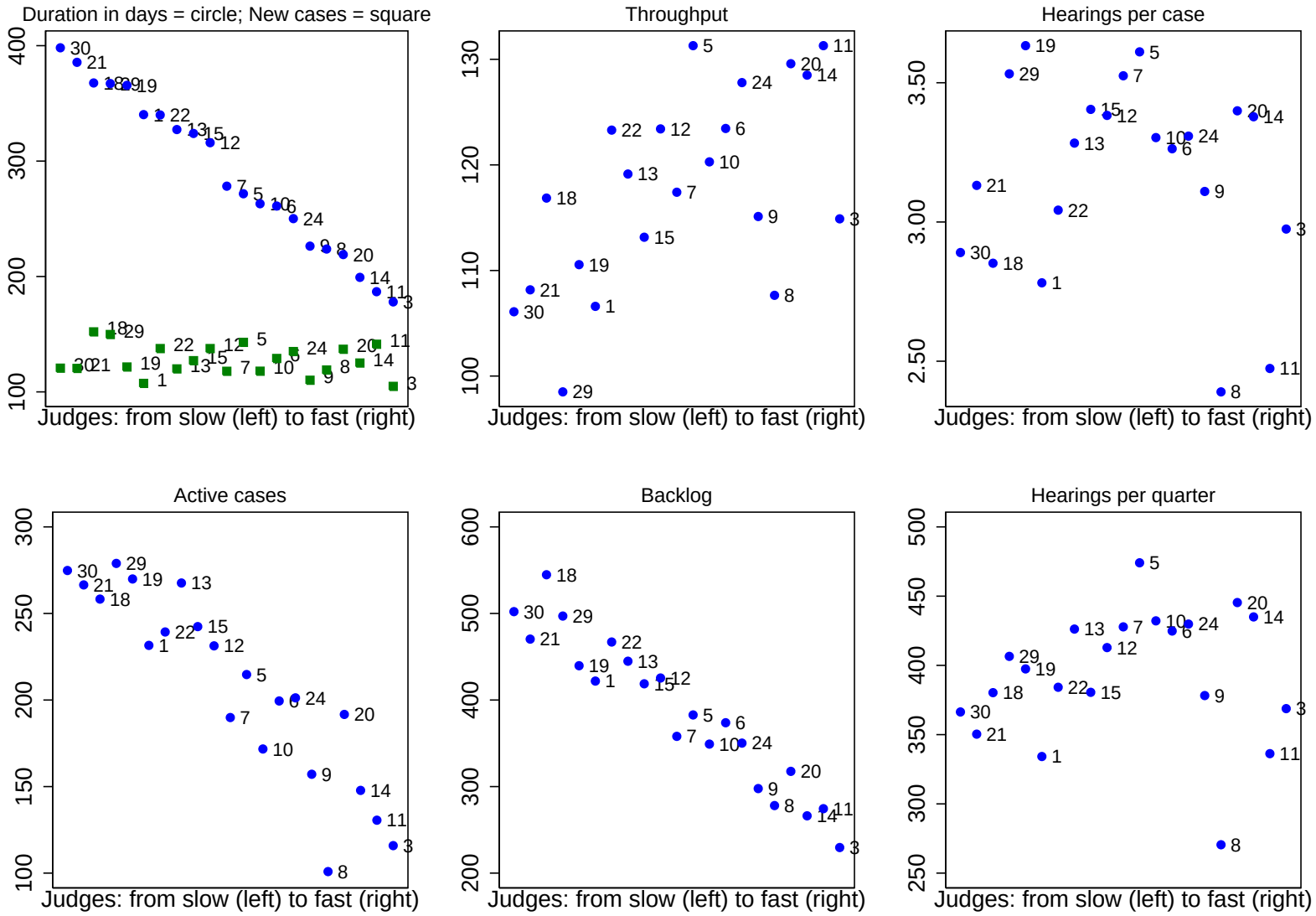

Figure 2: The trade off between quantity and quality in the decision of judges

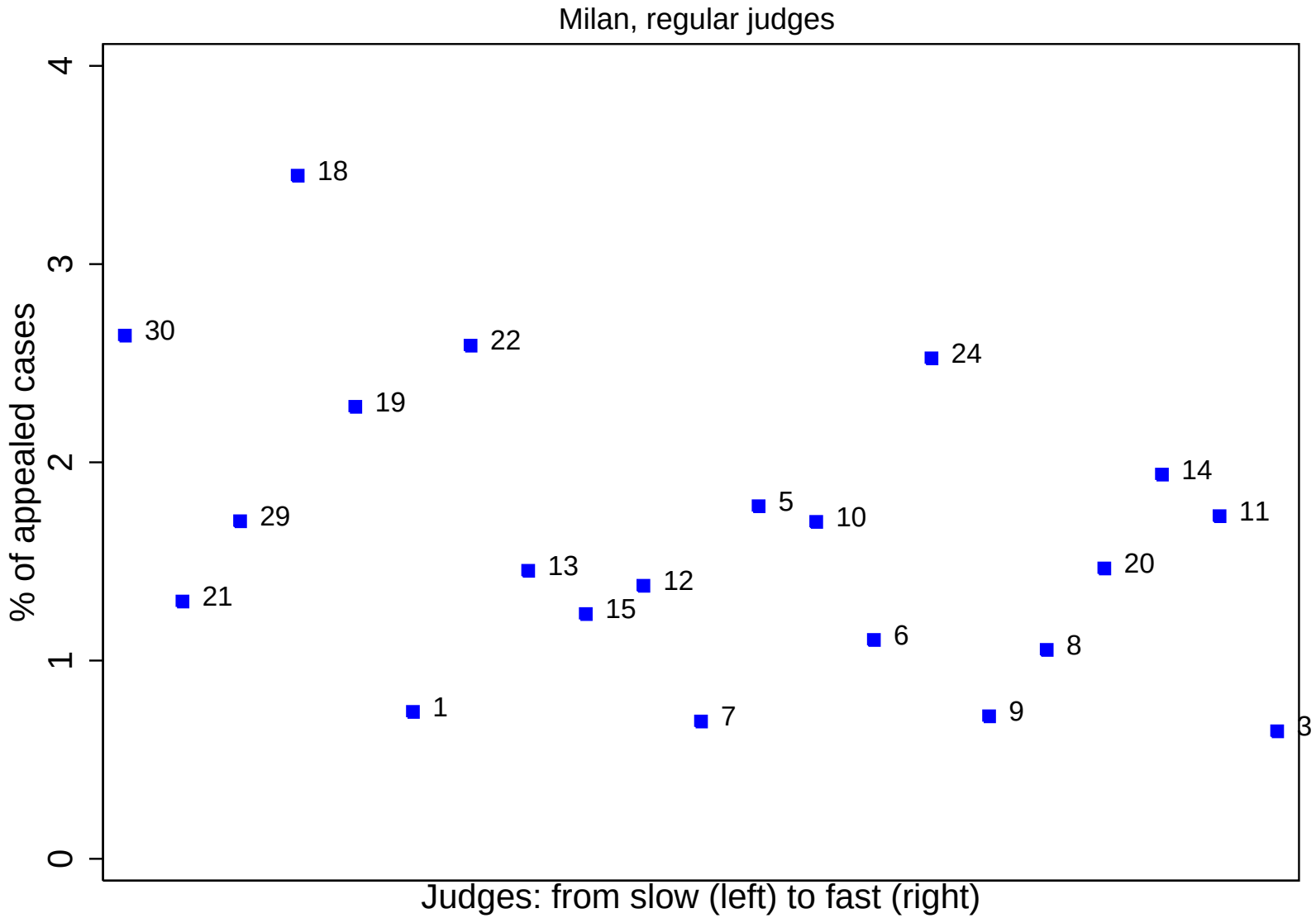

Figure 3: Work flow in a stable rotation

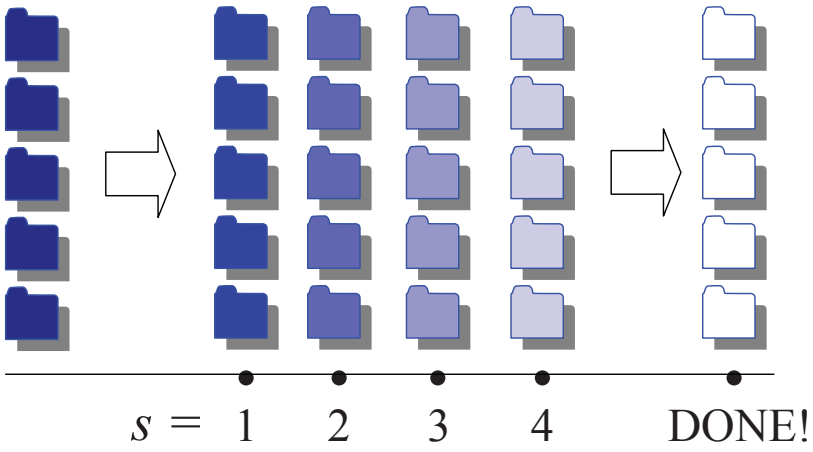

Figure 4: How far are judges from a stable rotation?

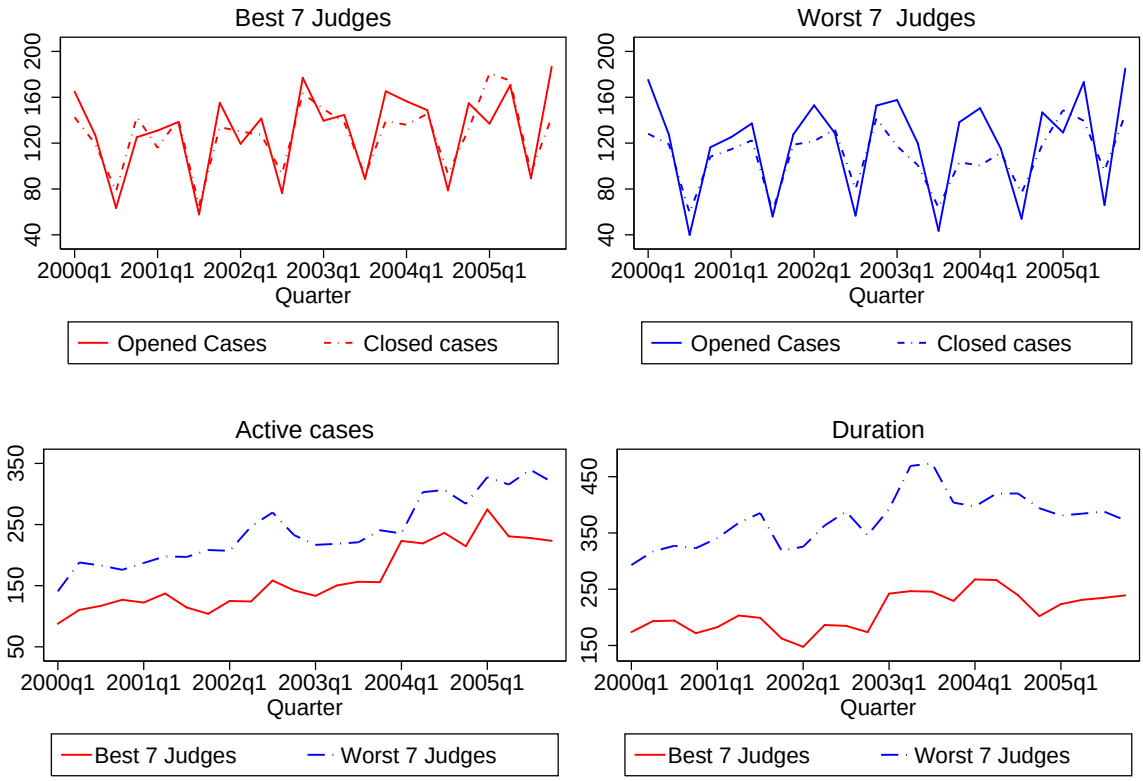

Figure 5: Deviation from a stable rotation

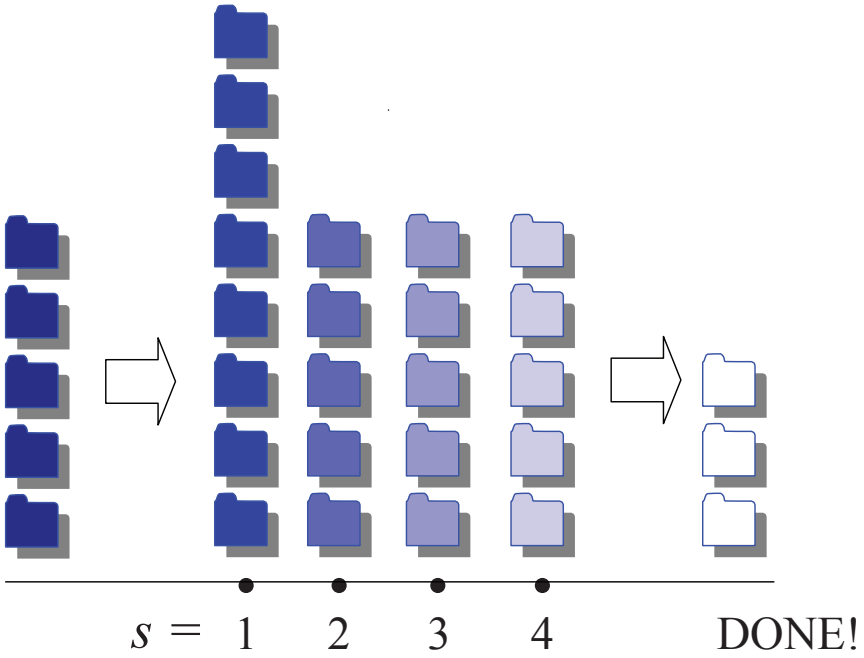

Table 3: Descriptive statistics

	Mean	sd	p25	p50	p75	n
Total duration	290	77	229	285	352	381
Inactive duration	126	49	91	120	155	381
Closed cases per quarter	119	35	98	122	145	381
Active cases at the end of quarter	210	74	154	206	266	381
New assigned cases per quarter	128	28	111	132	146	381
New urgent cases assigned per quarter	21	11	11	21	30	381
New complicated cases assigned per quarter	16	10	7	13	25	381
New cases beyond the "60 days limit"	130	27	114	133	148	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
New opened cases per quarter	127	46	93	137	159	381

Note: 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 4: The effect of parallel working on duration out of a stable rotation

Estimation Method	OLS	IV	OLS	IV
Variables	(1)	(2)	(3)	(4)
$A_{i,q}$: Active cases at the end of quarter	0.29	0.66		
	(0.07)	(0.30)		
$\nu_{i,q}$: New opened cases per quarter			0.39	0.79
			(0.08)	(0.33)
$\alpha_{i,q}$: New assigned cases per quarter	0.35	0.15	0.31	0.09
	(0.09)	(0.16)	(0.10)	(0.18)
$\frac{e}{S}_{i,q}$: Standardized effort per quarter	-0.67	-1.29	-0.84	-1.73
	(0.08)	(0.31)	(0.09)	(0.36)
Implicit trend	1.95	1.12	4.25	6.48
	(0.58)	(1.37)	(0.60)	(1.10)
F-first excluded instruments *(Joint)*		8.44		8.19
Sargan test *(p-value)*		0.67		0.74
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	0.55	0.38	0.54	0.37
R^2 including judges' fixed effects	0.85	0.80	0.85	0.80

Note: Robust standard errors in parentheses. 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. F-first excluded instruments (Joint) denotes the minimum eigenvalue of the joint first-stage F-statistic. When denoted with "Yes", regressions include *Judges FE* (21 dummies); *Quarter dummies* (2000q1-2005q4).

Table 5: First stage out of a stable rotation

Endogenous dependent variable	$\frac{e}{S}$	A	ν
Variables	(1)	(2)	(3)
New urgent cases assigned per quarter	1.32	-0.08	0.80
	(0.31)	(0.47)	(0.26)
New complicate cases assigned per quarter	-0.93	0.10	-0.10
	(0.29)	(0.44)	(0.24)
Cases beyond the "60 days limit"	-0.01	0.55	0.44
	(0.13)	(0.18)	(0.12)
New assigned cases per quarter	-0.16	0.19	0.09
	(0.17)	(0.17)	(0.11)
F-first excluded instruments (Joint)		8.44	8.19
Judges fixed effects	Yes	Yes	Yes
Quarters fixed effects	Yes	Yes	Yes
Observations	381	381	381
Number of Judges	21	21	21
R^2	0.76	0.72	0.79
R^2 including judges' fixed effects	0.75	0.84	0.79

Note: Robust standard errors in parentheses. 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. F-first excluded instruments (Joint) denotes the minimum eigenvalue of the joint first-stage F-statistic (matrix). When denoted with "Yes", regressions include *Judges FE* (21 dummies); *Quarter dummies* (2000q1-2005q4).