Kremer, Mirko; de Véricourt, Francis

Working Paper

Mismanaging diagnostic accuracy under congestion

ESMT Working Paper, No. 22-01

Provided in Cooperation with:
ESMT European School of Management and Technology, Berlin

Suggested Citation: Kremer, Mirko; de Véricourt, Francis (2022) : Mismanaging diagnostic accuracy under congestion, ESMT Working Paper, No. 22-01, European School of Management and Technology (ESMT), Berlin, https://nbn-resolving.de/urn:nbn:de:101:1-20220330125206660136 77

This Version is available at:
https://hdl.handle.net/10419/251899

ISSN 1866-3494

March 29, 2022

ESMT Working Paper 22-01

Mismanaging diagnostic accuracy under congestion

Mirko Kremer, Frankfurt School of Finance and Management

Francis de Véricourt, ESMT Berlin

Submitted to *Operations Research*
manuscript OPRE-2020-06-353.R2

Mismanaging Diagnostic Accuracy under Congestion

Mirko Kremer
Frankfurt School of Finance and Management, 60322 Frankfurt am Main, m.kremer@fs.de

Francis de Véricourt
ESMT European School of Management and Technology, Schlossplatz 1, 10178 Berlin, francis.devericourt@esmt.org

To study the effect of congestion on the fundamental trade-off between diagnostic accuracy and speed, we empirically test the predictions of a formal sequential testing model in a setting where the gathering of additional information can improve diagnostic accuracy, but may also take time and increase congestion as a result. The efficient management of such systems requires a careful balance of congestion-sensitive stopping rules. These include diagnoses made based on very little or no diagnostic information, and the stopping of diagnostic processes while waiting for information. We test these rules under controlled laboratory conditions, and link the observed biases to system dynamics and performance. Our data shows that decision makers (DMs) stop diagnostic processes too quickly at low congestion levels where information acquisition is relatively cheap. But they fail to stop quickly enough when increasing congestion requires the DM to diagnose without testing, or diagnose while waiting for test results. Essentially, DMs are insufficiently sensitive to congestion. As a result of these behavioral patterns, DMs manage the system with both lower-than-optimal diagnostic accuracy and higher-than-optimal congestion cost, underperforming on both sides of the accuracy/speed trade-off.

Key words: Congestion, Diagnostic accuracy, Experiments, Partially Observable Markov Decision Process, Path-dependent Decision Making, Undertesting, Task Completion Bias

1. Introduction

The management of diagnostic processes drives efficiency and quality in many manufacturing and service settings. At the heart of most diagnostic processes is the search for, and the assessment of, information. Such processes are difficult to manage because they require the decision maker (DM) to dynamically balance the benefit of acquiring more diagnostic information against the cost of doing so. When additional and unattended diagnostic tasks build up over time, making this trade-off becomes especially challenging. Yet, congestion is pervasive in diagnostic and search problems, including health care systems (Alizamir et al. 2013, 2019) support centers and help desks (de Vericourt and Zhoug 2005), and research project pipelines (Loch and Terwiesch 1999) such as drug development (Girotra et al. 2007). Congestion creates time pressure in the form of accumulation that endogenizes the cost of acquiring more information. To account for these

dynamic search costs, the DM needs to adjust the strategy they would apply in the absence of congestion.

Such congestion-adjusted behavior has been the focus of a growing body of empirical studies. Following early work of Schultz et al. (1998, 2003), this literature documents congestion-dependent decision making and its various quality-related short- and long-term implications (Delasay et al. 2019). For instance, DMs respond to increasing congestion by working faster (Staats and Gino 2012), or changing the order of tasks (Ibanez et al. 2018). Perhaps most critical to quality implications, DMs reduce work content as system congestion increases (Batt and Terwiesch 2016). For example, in response to increasing system congestion, DMs may respond by reducing testing or eliminating it completely. While such reduction of work content appears troublesome, sacrificing short-term accuracy in order to manage future congestion costs might well be desirable for the overall system performance. In other words, a DM that begins to "cut corners" when congestion is rising may well do so on normative grounds. In fact, the DM may not cut corners enough.

This raises important questions regarding congestion-dependent behavior. First, is behavior appropriately sensitive to system congestion, judged against some reasonable normative standard? Second, if bias exists, what are the underlying behavioral mechanisms? Third, if bias exists, does it have a first-order effect on system performance?

The goal of this paper is to offer initial answers to these questions, which poses the challenging task of characterizing managerial bias. Because this is particularly difficult in field settings where the complicated dynamics of congestion-prone environments effectively prevent the observer from characterizing optimal behavior with the same precision usually provided by theory, we instead study behavior under controlled laboratory conditions. This allows for a direct test of rigorous decision theory that provides a sensible benchmark and makes explicit the links to the dynamics and performance of the system at large (Allon and Kremer 2019).

We study load-dependent decision making in the context of diagnostic processes in capacitated systems with discretionary task completion. The DM faces a stream of diagnostic tasks. Each task consists of an elementary search problem, in which the DM sequentially runs imperfect tests to determine whether or not an item is faulty. At any time instant, the DM needs to decide whether to continue the search for additional diagnostic information or to stop and make a diagnosis. The DM faces an accuracy/congestion trade-off. The DM incurs a cost for misdiagnosing the item, which provides an incentive to run additional tests. But running a test takes time and new tasks might accumulate until they are attended to, which generates congestion costs. The main implication is that, in contrast to related search settings without congestion (Busemeyer and Rapoport 1988, Saad and Russo 1996), stopping thresholds may dynamically adjust to system congestion.

To assess whether behavior in this setting is appropriately sensitive to varying system congestion, we first present a sequential hypothesis testing model that provides a normative benchmark. Building on Alizamir et al. (2013), we characterize the structure of the optimal control policy for our set-up. Optimally managing the trade-off between accuracy and congestion requires the careful balance of three congestion-dependent stopping rules. Specifically, and as notation mnemonic, the DM can stop the diagnostic process: Ⓐ *after* receiving a test result; Ⓑ *before* testing at all; Ⓦ *while* waiting for a test result. Notably, rule Ⓐ is at the heart of search in settings without congestion, but the other two rarely (Rule Ⓑ) or never (Rule Ⓦ) are. To the best of our knowledge, we are the first to study how the latter two rules depend on congestion and affect system performance.

We develop behavioral hypotheses around these rules, including their sensitivity to varying system congestion, as well as the effect of state-dependent individual-level behavior on system-level dynamics and performance. We present the results from a series of experiments designed to test these hypotheses. Our data provide evidence that diagnostic testing behavior is *insufficiently sensitive to congestion*. We observe substantial undertesting at low congestion levels, where running diagnostic tests is cheap. DMs tend to stop after the first test result (Rule Ⓐ), and often stop before testing at all (Rule Ⓑ). While the resulting undertesting at low congestion levels should keep congestion low (at the expense of low diagnostic accuracy), this is not what we observe. Instead, DMs allow the system to reach higher levels of congestion, and remain at these levels for longer than predicted under the optimal policy. DMs achieve these high congestion levels more often than predicted because they do not abort enough while waiting for a test result (Rule Ⓦ) when new tasks start accumulating. DMs remain at these high congestion levels longer than predicted, because they do not diagnose enough without testing (Rule Ⓑ) at these levels.

Interestingly, these behaviors result in lower-than-optimal diagnostic accuracy despite higher-than-optimal congestion-related search cost. This pattern is in sharp contrast to search settings without task accumulation, for which the literature on deferred decision making and sequential hypothesis testing (Busemeyer and Rapoport 1988, Saad and Russo 1996) provides robust evidence that DMs search too little at the expense of decision quality (Palley and Kremer 2014). Seemingly, the presence of congestion is detrimental to both sides of the accuracy/search cost trade-off, due to decision biases that have little or no bearing in settings without congestion.

We present four sets of experiments designed to tease apart behavioral mechanisms underlying the observed performance losses, and test managerial levers to debias behavior. Study 1 provides the baseline, and evidence that the main observed choice and performance patterns are robust to whether the DM operates in an environment with low or high *ex ante* (i.e., prior to diagnostic testing) uncertainty about the item type. Study 2 provides the DMs with statistical support, to test the idea that performance loss is due to DMs' inability to correctly update their beliefs based on

the diagnostic evidence they have collected thus far. Study 3 removes congestion and related cost from the system, to test the idea that DMs might undertest at low congestion levels because they overestimate future congestion-related costs of testing. Study 4 decomposes two possible sources of performance losses, and provides evidence that the gap with the optimal policy is equally due to the DMs' poor choice of strategy and the poor execution of this strategy. In fact, forcing DMs to implement their own strategies, even poor ones, improves performance.

2. Literature

Our study is related to a growing empirical literature that shows how human servers adjust processing times in response to increasing system load (see Delasay et al. 2019 for a comprehensive review). For example, workers tend to accelerate their work rate under high workload conditions, in task contexts that include toll booths at bridges (Edie 1954), serial production lines (Schultz et al. 1998), grocery retail (Mas and Moretti 2009, Lu et al. 2013, Wang and Zhou 2018), intra-hospital transport (Kc 2009), loan-processing (Staats and Gino 2012), emergency departments (Batt and Terwiesch 2016), and restaurant service (Tan and Netessine 2014).

In contrast, we study a setting in which servers have the discretion to respond to growing congestion levels simply by reducing the number of tasks they execute, rather than speeding up their execution. For example, Oliva and Sterman (2001) find that bank-office workers spend less time processing loan applications when congestion increases. Powell et al. (2012), in the context of hospital reimbursement processes, find that physicians reduce the diligence of paperwork execution as response to increasing workload. Kc and Terwiesch (2012) study the load-dependent rationing of bed capacity of cardiac intensive care units, and find that patients are discharged earlier in "busy" ICUs. A number of follow-up studies provide further evidence on load-dependent task reduction (Kuntz et al. 2015, Jaeker and Tucker 2017), although some studies do not find evidence for load-dependent patient discharge (Keenan et al. 1998, Kim et al. 2015, Chan et al. 2019).

While speculative, part of this mixed pattern might be that these studies rely on patients' length-of-stay (LOS) as a measure of work content. Aggregated across various activities that are embedded in multi-stage and parallel processes operated by various resources, LOS includes less critical activities as well as non-value-adding activities such as waiting. Indeed, in the context of a maternity unit, Freeman et al. (2017) find that general practitioners respond to increasing workload by cutting activities that are not central to the primary service outcome. Similarly, Long and Mathews (2018) find LOS decreases in occupancy through a reduction in less critical boarding time, rather than through a reduction of critical care elements. This points to the importance of carefully opening up the construct work content, to understand what precisely it means to reduce it, and ideally measure it at a less aggregate level. Similar to our study, Batt and Terwiesch (2016)

directly measure "task reduction" by using the number of diagnostic tests ordered for a patient by a physician as a proxy for the work content of a patient. The authors find no evidence of doctors ordering fewer tests when the system is busy, which aligns qualitatively with the non-results reported in Forster et al. (2003) and Mæstad et al. (2010). Overall, there seems to be consensus in the literature that task reduction negatively affects the quality of service, via various direct and indirect mechanisms. On the other hand, the literature is inconclusive on whether congestion-dependent task reduction exists, and largely silent on whether it should.

Our study contributes to this literature for a number of reasons that relate to our ability to control and manipulate essential aspects of the task environment experimentally (rather than econometrically), such as arrival and service rates, population base rates, and even whether or not tasks accumulate. While the standard arguments regarding external validity apply, the controlled setting of our study has two main advantages. First, it allows to tease apart reasons for why DMs may or may not engage in task reduction. Second, it allows for a rigorous benchmark against which to assess observed behavior as biased or not. Despite such advantages, experimental research on load-dependent server behavior remains scarce (see Hathaway et al. 2021 for a recent exception). Our study is an attempt to fill this gap.

3. Task Setting: Model and Theory

Our goal is to study empirically the sensitivity of diagnostic decisions to system congestion, in a setting where system congestion itself is (partially) determined by decisions. We want to study this in a task setting that allows us to qualify whether observed behavior is appropriately sensitive to system congestion, relative to a reasonable normative benchmark. We next present such a task setting, and characterize optimal decision making for it.

3.1. Basic Model

We consider a discrete time version of the task setting in Alizamir et al. (2013). In this set-up, a DM runs a diagnosis on items that arrive sequentially to the system. Items are processed on a first-come-first-served basis and accumulate when not attended to. The DM can conduct sequential tests on an item to determine whether or not it is faulty. Items are identical a priori with a prior probability of being faulty equal to p_0.

Discrete time. In our set-up, a time period corresponds to one unit of time. At the end of each period, either a new item arrives with probability λ or a test result is obtained with probability $1-\lambda$ (if a test was ordered). This system is equivalent to a continuous time process, where items arrive according to a Poisson process and each test takes an exponentially distributed time (see

Bertsekas 2005a). In the following, we refer to time period t as time interval $[t, t+1)$, which is the time between time epochs t and $t+1$, with a slight abuse of notation.

Diagnostic tests. The DM can order tests to diagnose an item. Each test either succeeds or fails. A faulty item always fails the test, and thus a successful test always reveals that the item is not faulty. By contrast, a non-faulty item successfully passes the test only with probability $\beta \in (0.5, 1]$. We denote by p_k the probability that an item is faulty after having failed $k > 0$ tests, and we have from Bayes' rule $p_k = p_{k-1}/(1-\beta+\beta p_{k-1})$, which increases in k.

Diagnostic costs. At any time epoch, the DM can stop the diagnostic process. When stopping the process, the DM incurs penalty ρ if she makes the wrong diagnosis and the expected cost of misdiagnosing an item after k failed tests is equal to $\rho \times (1-p_k)$. However, items may accumulate while the DM is busy running tests. The congestion cost for one period is equal to cx, where c is the cost per item and per period and x the total number of items in the system.

Decisions and dynamics. At any time epoch t, the DM can either i) release the current item by making a diagnosis and move to the next one, or ii) conduct a test on the current item. The system state is given by (x, k). If a new item arrives, the system moves from state (x, k) to state $(x+1, k)$. If a test is completed (with $x > 0$), the system moves to state $(x, k+1)$. And if the DM decides to make a diagnosis for and release the current item, the system moves to state $(x-1, 0)$ and the DM starts working on a new item. Because faulty items always fail the test, it is always optimal to diagnose the item as not faulty as soon as an item passes a test. Otherwise, the DM needs to decide whether to stop (option i) or continue the current diagnostic process (option ii).

Accuracy/congestion trade-off. The DM faces a fundamental trade-off between accuracy and congestion: The more she spends time diagnosing an item, the lower the expected misdiagnosis penalty, but the more items may accumulate in the system and thus the higher the congestion costs. More formally, the DM's actions at any time epoch defines their policy $u(\cdot)$, such that $u(t) = i$ and $u(t) = ii$ if the DM chooses option i) and ii) at time epoch t, respectively. The performance of a policy is measured as the long-run average congestion and misdiagnosis cost per diagnosed item,

$$g^u = \liminf_{T\to\infty} \frac{1}{\lambda T} \mathrm{E}\left[\rho \bar{M}^u(T) + \rho \underline{M}^u(T) + \sum_{t=0}^{T} cX^u(t)\right] \tag{1}$$

where $\bar{M}^u(t)$ (resp. $\underline{M}^u(t)$) is the random cumulative number of misidentified faulty (resp. non-faulty) items up to t; and $X^u(t)$ is the random number of items in the system at t.

In essence, the DM faces a sequential testing problem. Without congestion, this corresponds to the elementary stopping time problem (Bertsekas 2005b), where the cost of running a test equals c. With congestion, this cost is equal to cx and changes dynamically with the arrival of new items. Thus, a key feature of our set-up is that the DM can stop the search (option i) under three very distinct conditions, i.e., either *after* receiving a test (Rule Ⓐ), *while* waiting for a test's result (Rule Ⓦ), and possibly even *before* ordering any test (Rule Ⓑ).

3.2. Structure of the optimal decision rule

The problem of finding the optimal decision rule which minimizes (1) corresponds to a Partially Observable Markov Decision Process. Alizamir et al. (2013) fully characterize this optimal rule and the following propositions state their results in our set-up.

Proposition 1a (Diagnosis). The DM diagnoses an item as not faulty if it passes a test. Else, she diagnoses the item as faulty if $p_k \geq 0.5$ and as not faulty otherwise, when she stops the process.

Proposition 1b (Stopping Decision). A threshold $\hat{p}$ on p_0 exists such that

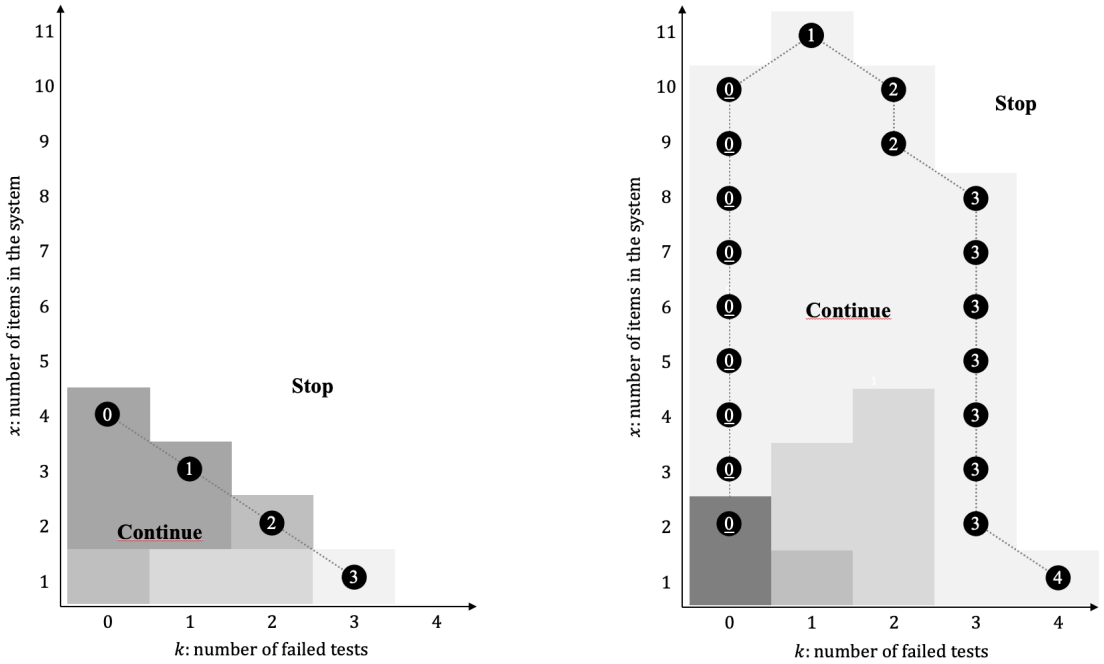
- Case $p_0 \geq \hat{p}$: thresholds $\bar{k}(x)$ exist such that in state (x,k), the DM i) runs an additional test if $k < \bar{k}(x)$, and ii) stops otherwise. Further, threshold $\bar{k}(x)$ is non-increasing in congestion level x.
- Case $p_0 < \hat{p}$: thresholds $\underline{k}(x)$ and $\bar{k}(x)$ exist such that in state (x,k), the DM i) runs an additional test if $\underline{k}(x) < k < \bar{k}(x)$, and ii) stops otherwise. Further, thresholds $\underline{k}(x)$ and $\bar{k}(x)$ are non-decreasing and non-increasing, respectively, in congestion level x.

"High uncertainty" environments. Figure 1a summarizes the key predictions for $p_0 \geq \hat{p}$. In essence, the proposition states that the DM should continue to diagnose the current item as long as the number of performed tests is below a stopping threshold $\bar{k}(x)$, which decreases in congestion level x. When congestion grows, the DM diagnoses with less information via one of two stopping rules. First, they stop after receiving a test (Rule Ⓐ), and do so earlier at higher levels of congestion. Second, they stop while waiting for a test (Rule Ⓦ) as congestion grows. Note that making a diagnosis without running any test (Rule Ⓑ) is never optimal in steady state for $p_0 \geq \hat{p}$.

"Low uncertainty" environments. Figure 1b summarizes the structure of the optimal stopping policy when $p_0 < \hat{p}$. The main difference for decision making in this environment is that the DM faces relatively low diagnostic uncertainty at p_0, such that failed tests may actually *increase* uncertainty early in the diagnostic process.[1] As a result, the optimal policy at medium congestion levels is conditional on how far along the diagnostic process is. Using $x = 4$ for illustration, the DM should stop at $\underline{k}(4) = 0$ because of reasonably good diagnostic odds ($p_0 = 0.14$), continue testing at $k = 1$ and $k = 2$ because of relatively bad odds ($p_1 = 0.4$, $p_2 = 0.73$), and stop at $\bar{k}(4) = 3$ because of reasonably good odds ($p_3 = 0.98$). Importantly, while DM may occasionally sustain higher congestion levels to reduce the increased diagnostic uncertainty after the first and second failed test, she subsequently limits congestion by diagnosing items as non-faulty without testing at $\underline{k}(x) = 0$ - using Rule Ⓑ is optimal in steady state for $p_0 < \hat{p}$.

[1] Indeed, for the parameters of Figure 1b, p_k is increasing in k such that $p_0 < p_1 = 0.4$ and hence p_1 is closer to 0.5 than p_0 making the decision more uncertain.

Figure 1 Optimal policy for the parameters of our studies: $\beta = 0.75$, $\lambda = 0.5$, $c = \$1$, $\rho = \$100$.



(a) *High* uncertainty ($p_0 = 0.40$) **(b)** *Low* uncertainty ($p_0 = 0.14$)

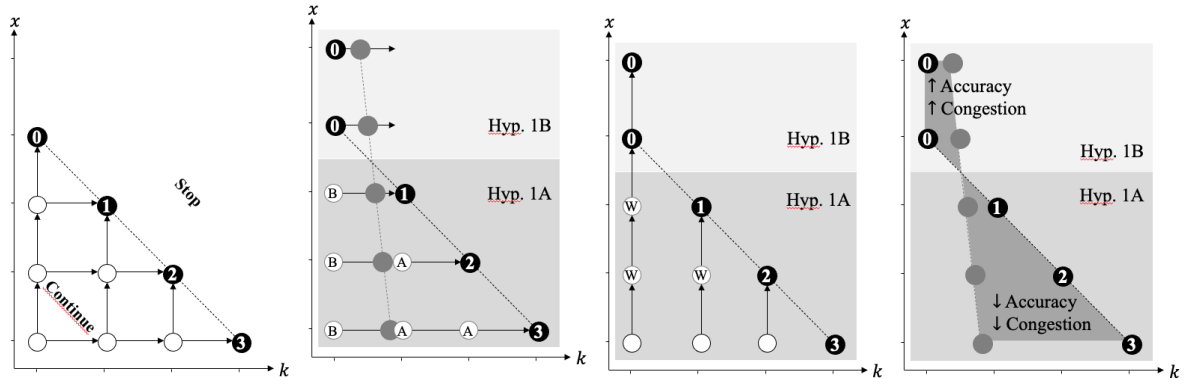
Note: Thresholds $\bar{k}$ and $\underline{k}$ indicate $\bar{k}(x)$ and $\underline{k}(x)$, respectively; the grey areas represent the frequency (in %) at which the system visits a state under the optimal policy with: $< 1\%$ 1-5% 5-10% 10-15% $> 15\%$.

Path-dependence. The current congestion (x) and number of failed tests (k) fully summarize the necessary information to make a decision. Thus, optimal behavior shows no path-dependence, i.e., decisions only depend on state (x, k), regardless of how the system arrived in that state.

4. Hypotheses and Implementation

Against the backdrop of the task environment described in the previous section, our objective is two-fold. First, we want to understand human stopping behavior, relative to theoretical predictions. Whether DMs stop diagnostic processes too late or too early, and whether stopping behavior is appropriately sensitive to congestion, essentially depends on DMs' biases relative to the use of the three stopping rules (Ⓑefore, Ⓐfter, Ⓦhile) under the optimal policy. Second, we want to study the link between such individual-level behavior and system-level dynamics and performance. Assessing such a link is empirically challenging, because the decisions made in each state influence the system dynamics, which in turn affect future decisions. Towards developing empirically testable hypotheses we next decompose this difficult prediction problem into smaller, manageable, pieces. Figure 2 provides a summary and preview of this exercise.

Figure 2 Predictions (●) and Data (●): behavioral mechanisms and hypotheses (illustration for $p_0 = 0.4$)



(a) Theory: Stopping thresholds $\bar{k}$

(b) Hyp 1: Under/Overtesting
Hyp 2: Sensitivity

(c) Hyp. 3 and Rule Ⓦ: Stop while waiting

(d) Hyp 4: Cost predictions

4.1. Hypotheses

4.1.1. Stopping at low congestion levels. We first focus on stopping behavior in "uninterrupted" diagnostic processes, i.e. sequences of failed test results that are *not* interrupted by intermittent arrivals. These sequences do not experience an increase in congestion while testing. As such, they provide a clean mapping to sequential hypothesis testing settings without congestion.

Given this, we predict that DMs undertest at congestion levels where testing is relatively inexpensive (such that $\bar{k}(x) > 0$). This prediction is broadly consistent with empirical early-stopping results in related search settings *without* congestion, but the underlying reasons are more multifaceted in our relatively richer setting *with* congestion. To anchor the arguments, note that the stopping decision in our context is akin to the choice between two lotteries. Stopping represents a simple gamble between losing penalty ρ (with probability $\mu = \min(p_k, 1-p_k)$[2]) and losing nothing (with probability $1-\mu$)). In contrast, the decision to continue is a complex gamble, characterized by the expected costs from future decisions as well as the system dynamics.

Early stopping may then arise if DMs simply do not align with the preferences that underlie the prediction of our benchmark model. It stands to reason that DMs have a preference for the simpler and less ambiguous stopping gamble. As a result, DMs may willingly stop the diagnostic process at probability $p_{k(x)} < p_{\bar{k}(x)}$, even in the absence of any judgment bias with regard to $p_{k(x)}$. An extreme case would be the decision to stop the process Ⓑefore ordering a test.

Rule Ⓑ. DMs stop before testing, at $k = 0$.

[2] Indeed, when stopping, the DM should declare the the item as faulty if $p_k > 1-p_k$ (and as not faulty otherwise), in which case the DM is correct with probability p_k. Thus, the DM makes a mistake with probability $1-p_k$ when $p_k > 1-p_k$, and with probability p_k when $p_k \geq 1-p_k$, i.e. $\mu = \min(p_k, 1-p_k)$.

But even if DMs do test ($k > 0$), they may stop too early Ⓐfter receiving the result from test $k < \bar{k}(x)$. The judgment literature provides plenty of reasons for the DM to hold flawed posterior beliefs, miscalculating p_k, or misinterpreting p_k when it is known. Importantly, DMs tend to infer too much from small samples (of diagnostic tests, in our case), resulting in the relative underweighing of base rate probabilities and overreaction to signals (Kahneman and Tversky 1973, Rabin 2002). The resulting excessive adjustments of posterior beliefs in the direction of a test result may then lead DMs to stop too early, relative to the threshold $\bar{k}(x)$ defined in Proposition 1b.

Rule Ⓐ. DMs stop after receiving the k^{th} test such that $0 < k < \bar{k}(x)$.

Early stopping may also result when a DM mistakenly stops a diagnostic process simply due to randomness in decision-making, a notion that has recently been popularized in the operations literature (Su 2008, Huang et al. 2013). Indeed, Bearden and Murphy (2007) demonstrate how unbiased random error in stopping decisions can lead to systematically biased early stopping behavior in search problems without congestion. Remaining agnostic (for now) about the precise underlying reasons, rules Ⓐ and Ⓑ lead us to expect behavior qualitatively similar to the early stopping patterns from the search literature (Pitz et al. 1969). With reference to Figure 2b, we hypothesize:

Hypothesis 1A. (● < ●) DMs order and receive fewer-than-optimal tests per diagnosis on average, at each level of congestion x for which $\bar{k}(x) > 0$.

4.1.2. Stopping at high congestion levels. As our ultimate interest is in stopping behavior on average, across congestion levels, we need to predict decision making at high congestion levels where $\bar{k}(x) = 0$. Because undertesting is not possible when $\bar{k}(x) = 0$, we hypothesize that DMs occasionally stop too late. For example, the same notion of random error that may induce early stopping when testing is optimal, can lead to late stopping when theory predicts no testing at all. With reference to Figure 2b, we hypothesize:

Hypothesis 1B. (❶ < ●) DMs order and receive more-than-optimal tests per diagnosis on average, at each level of congestion x for which $\bar{k}(x) = 0$ or $\underline{k}(x) = 0$.

Although the under/overtesting pattern of Hypotheses 1A and 1B is compatible with stopping decisions that are entirely insensitive to congestion, we expect these decisions to reflect that the cost of improving diagnostic accuracy increases with congestion. This prediction aligns with the empirical stopping literature which documents that DMs acquire less information as the cost of doing so increases (Rapoport and Tversky 1970). The main difference is that the expected search cost in our setting derives from current and future congestion levels, i.e. the search cost is uncertain and changes dynamically. We thus expect some sensitivity of behavior to congestion, consistent

with the empirical literature on load-dependent service times (Delasay et al. 2019). Our theoretical model allows us to qualify whether stopping behavior is appropriately sensitive - Hypotheses 1A and 1B imply that it is not (Figure 2b).

HYPOTHESIS 2. (vs.) DMs order and receive fewer tests at higher congestion levels - stopping behavior is sensitive to congestion, but insufficiently so.

4.1.3. Stopping when congestion grows. Hypotheses 1 and 2 predict stopping behavior at different congestion levels, but remain silent on behavior as the system transitions between congestion levels. A key property of our task environment is that DMs may need to stop a diagnostic process when congestion grows while they are waiting to receive a test result (Figure 2c).

RULE Ⓦ. DMs stop while waiting for test results.

The key question regarding this stopping rule is: are DMs doing too much of it, or too little? We ground our expectation regarding Rule Ⓦ on the idea that stopping decisions in a given system state (x, k) are independent of how the system had transitioned into that state. Specifically, if the DM stops after receiving a test result (Rule Ⓐ: $(x, k-1) \rightarrow (x, k)$), she should also stop after *not* receiving a test result (Rule Ⓦ: $(x-1, k) \rightarrow (x, k)$). This principle of path-independent behavior is normatively compelling, and provides a sensible baseline prediction for behavior.

HYPOTHESIS 3. For any state (x, k), the stopping decision is path-independent.

Although silent on whether DMs stop too much or too little while waiting for test results, the hypothesis ties together different stopping rules. Specifically, if the DM stops too much after receiving a test result (Rule Ⓐ), Hypothesis 3 implies that the DM also stops too much after *not* receiving a test (Rule Ⓦ), all else being equal (i.e., for the same system state).

Of course, the behavioral literature provides several results that would challenge Hypothesis 3 on the idea that stopping behavior is equally sensitive to additional diagnostic information as it is to additional congestion, based on the notion of psychic wait costs that include psychological sensations such as anxiety or stress. In contrast to the standard assumptions made in formal queuing models (including ours), psychic wait cost need not be linear, or even monotonically increasing, in wait time (see Allon and Kremer 2019 for a review of the related literature). On the one hand, if psychic wait cost increase excessively while waiting for a test result, DMs might stop a diagnostic process earlier than theoretically predicted. On the other hand, because ordering a test triggers a wait towards the goal of receiving the result, goal theory would predict that the wait cost might decrease as the DMs gets closer to the goal (Kivetz et al. 2006). In this case, DMs might stop diagnostic processes later than theoretically predicted. Both psychic forces are behaviorally plausible and could even coincide (Janakiraman et al. 2011). Whether and how they translate to the specific task context of our study is a question that we try to answer empirically.

4.1.4. System-level behavior. Given the hypothesized stopping behavior, and how it is moderated by varying congestion levels, what can we expect in terms of congestion cost and diagnostic cost, relative to optimal decision making? Figure 2d illustrates that the answer is not clear a priori, and requires further thought on how the system is likely to behave in light of the stopping behavior we predict. On the one hand, for states that require testing, Hypothesis 1A predicts that DMs undertest and hence enjoy lower-than-optimal congestion costs at the expense of higher-than-optimal diagnostic costs. On the other hand, for states that require no testing (including states that the optimal policy never reaches), Hypothesis 1B predicts that DMs overtest and hence enjoy lower-than-optimal diagnostic costs at the expense of higher-than-optimal congestion costs.

Predicting system performance then depends on the relative strength of over- versus undertesting bias in the two types of states, as well as the relative occurrence of these states, which is linked to system dynamics. While precise predictions are elusive, our Hypotheses allow for cautious extrapolation regarding system behavior. Specifically, we predict undertesting at low congestion levels (Hyp. 1A), which in itself would tend to keep congestion low. Whether the arrivals of new items might increase congestion to levels where we predict overtesting (Hyp. 1B and 2) then depends on the DM's propensity to stop when she sees congestion grow while waiting for a test result (Rule Ⓦ). When DMs at low congestion levels stop quickly after a test return, the assumed path-independence (Hyp. 3) predicts that DMs stop equally quickly when they see an increase in congestion. This limits substantially the number of paths via which the system can reach high congestion levels. Overall, we expect DMs to be more likely to enter states where undertesting leads to excessive diagnostic penalty cost, than states where overtesting leads to excessive congestion cost.

Hypothesis 4 (Cost). DMs incur lower-than-optimal congestion costs, at the expense of higher-than-optimal diagnostic penalty costs.

We briefly outline alternative cost patterns, and the likelihood of them arising in our data. First, no behavioral pattern can lead to both lower-than-optimal congestion cost and lower-than-optimal diagnostic cost. Second, DMs may enjoy lower-than-optimal diagnostic cost at the expense of higher-than-optimal congestion cost, if they were to generally overtest. Given existing evidence from the literature, we do not find such a pattern likely. Third, it is possible that DMs stop early after they receive test results (Rule Ⓐ) but continue the diagnostic process when they wait for a test result (Rule Ⓦ). Such behavior would effectively violate the path-independence property of the optimal policy (Hypothesis 3), and be consistent with the possibility that DMs incur higher-than-optimal diagnostic cost *and* higher-than-optimal congestion cost (violating Hypothesis 4).

4.2. Design and implementation (all studies)

We present the results from four studies designed to test our research hypotheses, including a number of tests of the behavioral mechanisms that may underlie the predicted decision patterns. This section presents the experimental design elements that are common to all studies.

Table 1 Roadmap to studies

Study	H1a	H1b	H2	H3	H4	Behavioral mechanism	Debiasing mechanism	Experimental factor(s)
1	✓	✓	✓	✓	✓	-	-	Diagnostic uncertainty p_0: *Low* (0.14) vs. *High* (0.4)
2	✓	✓	✓	✓	✓	Judgment: Biased beliefs p_k	Statistical support	Access to p_k: *Show* vs. *Hide*
3	✓	×	×	×	×	Overestimation of congestion cost	-	Task accumulation: *Congestion* vs. *No congestion*
4	○	○	○	○	○	Biased strategy Biased execution	Strategy definition Strategy commitment	Strategy elicitation: yes vs. no Strategy commitment: yes vs. no

Notes: ✓tested. ○ testable. × not testable

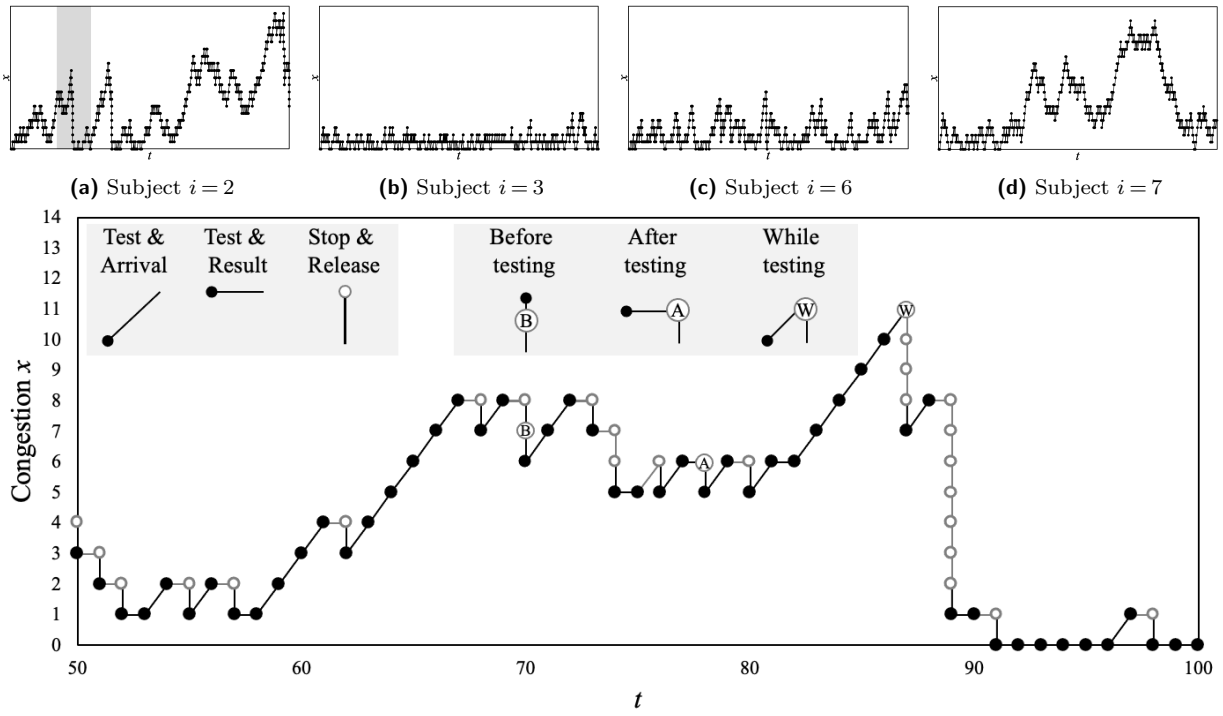
4.2.1. Task and Design. DMs in our experiments face the task described in Section 3 for $T = 400$ time epochs, with the parameters of the optimal policy in Figure 1. Thus, we set per-unit time cost $c = \$1$, test reliability $\beta = 0.75$, diagnostic penalty cost $\rho = \$100$, and arrival rate $\lambda = 0.5$, which means that a test return and a new arrival both have probability 0.5. Our experimental design varies prior p_0 (0.4 vs. 0.14, in Studies 1-3), whether DMs have access to the statistical posterior p_k (Studies 2-4) or not (Study 1), and whether the system is prone to congestion (Studies 1, 2, 4) or not (Study 3). All treatments are implemented in a between-subject design.

4.2.2. Prior Information and Sample Information. We provide subjects with full knowledge about the relevant parameters of the environment, which include financials (c and ρ) as well as stochastic elements (λ, p_0, β). To sharpen our statistical inferences, *all* experimental treatments use the same pre-generated sets of random realizations of item types, test results and arrival events. See Appendix EC.1.1 for more details.(All appendices are given in the e-companion to this paper.)

4.2.3. Data structure. Figure 3 illustrates the nature of our data, using a small sample from Study 2, plotting congestion level x over time epoch t. At the most granular level, for each subject i (we will suppress i in the following), the data comprises a sequence of states (x, k). The states are linked through exogenous events at time epoch t, and through decisions made over time period t (recall that time period t is defined as time interval $[t, t+1)$). An event (e_t) either corresponds to an arrival $(e_t = 1)$ that increases system congestion x_t without changing the number of failed tests k, or a test return $(e_t = 0)$ that changes k without changing congestion x_t. A decision corresponds

to either the continuation of the diagnostic testing process, which triggers the next exogenous event e_t or a diagnosis that reduces system congestion and resets $k=0$ for the next item in the queue.

Figure 3 Data and System Dynamics



(a) Subject $i=2$ **(b)** Subject $i=3$ **(c)** Subject $i=6$ **(d)** Subject $i=7$

(e) Subject i=2, $t=50-100$

Importantly, because the DM can diagnose multiple successive items without testing (Rule Ⓑ) between events e_t and e_{t+1}, we let y_τ denote the τ^{th} decision of the subject, and $s_\tau=(x_\tau,k_\tau)$ the corresponding system state. If the subject stops ($y_\tau=1$), she incurs a penalty of $\rho_\tau=\rho$ if the diagnosis is incorrect, and $\rho_\tau=0$ if it is correct. If the subject continues to test ($y_\tau=0$), which triggers the next event e_{t+1}, she incurs congestion cost cx_t, where x_t is the congestion level at the end of period t. Specifically, congestion level x_t associated with period t is equal to $x_{\tilde{\tau}}$ with $\tilde{\tau}=\max(\tau:t_\tau=t)$, and where t_τ is the time period during which the τ's decision occurs.

4.2.4. Performance metrics and aggregation. Our ultimate interest is in how decisions affect system performance. Towards this objective, we define the total number of diagnoses made until time epoch t as $D_t=\sum_{\tau:t_\tau\le t}y_\tau$. The number of tests ordered, O_t, and the number of tests received, R_t, are calculated in a similar way. Consistent with our theoretical model from Section 3, we calculate total congestion cost until time epoch t as $C_t=\sum_{z\le t}cx_z$, and total diagnostic penalty cost until time epoch t as $P_t=\sum_{\tau:t_\tau\le t}\rho_\tau$. We express all key metrics in *per diagnosis* terms: average congestion costs $\bar{C}=C_T/D_T$, average diagnostic costs $\bar{P}=P_T/D_T$, average number of tests ordered $\bar{O}=O_T/D_T$, and average number of tests received $\bar{R}=R_T/D_T$.

Defined at the subject level (we had dropped i), these metrics require further aggregation to the population level at which we wish to draw and report our conclusions. This requires some care, given that our experiments embrace the full dynamics of the system, which unfold as a function of exogenous stochastic events and each DM's own decisions. Essentially, our statistical tests rest on simple (as opposed to weighted) averages, to account for the fact that DMs generally contribute a different number of observations towards different population averages of interest.[3]

4.2.5. Benchmark. Our interest in assessing behavioral bias requires the comparison of our data with the "unbiased" predictions from Section 3. We can benchmark stopping behavior against optimal decision making on two levels. At the level of individual decisions, we compare average stopping (between 0 and 1) at every system state $s_\tau = (x_\tau, k_\tau)$ observed in our data, with the optimal policy (either 0 or 1). At the level of aggregate performance metrics ($\bar{C}$, $\bar{P}$, $\bar{O}$, $\bar{R}$), we calculate the corresponding benchmarks by applying the optimal policy on the same set of scenarios (i.e. realizations of the stochastic processes) as the ones we implement in our studies.

4.2.6. Software, Recruitment and Payment (Studies 1-3). We recruited subjects from an experimental subject pool associated with the Laboratory for Economic Management and Auctions (LEMA) at Pennsylvania State University. After arriving at the laboratory, participants read written instructions (see Appendices EC.1.2 and EC.1.3 for samples). The experiment was implemented in the software zTree (Fischbacher 2007). Subjects first played 20 time epochs to familiarize themselves with the software and the task. They then performed the task for $T = 400$ time epochs under incentive-compatible conditions. Subjects were told that the task would last between 350 and 450 time epochs, to mitigate possible (but unlikely) end game effects - note also that, given the incentive structure of our task setting, DMs cannot improve performance by changing their behavior towards T. Each session lasted about 45 minutes. Subjects were compensated based on the total cost averaged across all items diagnosed over the 400 time epochs. Specifically, subjects earn according to $B - v * (\bar{P} + \bar{C})$, where the fixed base payment B and the conversion rate v were set such that the earnings per hour of lab time were kept reasonably similar across experimental conditions. The average compensation was \$13 and participants were paid in private at the end of the session; cash was the only incentive offered.

[3] E.g., the number of diagnoses made (D_t) varies across DMs. Consider a simple illustrative example: DM A made $D_A = 1$ diagnosis based on $R_A = 2$ tests received. In contrast, DM B made $D_B = 2$ diagnoses, each based on a single test received (hence, $R_B = 2$). At the aggregate population level, we hence have a total of 3 (1+2) diagnoses made based on a total of 4 (2+1+1) tests, yielding an average of 1.33 ($=\frac{4}{3}$) tests received per diagnosis made. However, this calculation gives undue weight to DM B's lower number of tests per diagnosis, given that DM B made more diagnoses. Correcting for such a bias, the population-level average number of tests per diagnosis is in fact 1.5 ($=\frac{2+1}{2}$), i.e., the simple (*not* weighted by subject-level sample size) average of DMs' average number of tests received.

5. Study 1: Managing diagnostic processes and the role of p_0.

5.1. Design and Implementation

We implement two conditions that correspond to the two cases of Proposition 1b: $High$ (with $p_0 = 0.40$) and Low (with $p_0 = 0.14$). Recall that in $High$, uncertainty decreases as the diagnostic process progresses, i.e., p_k increases above and away from 0.5 in k. In contrast, uncertainty increases in the beginning of the process in Low, i.e., p_k increases towards 0.5 for small values of k.

DMs should benefit from being in a Low uncertainty environment. For the parameters of our study, theory predicts a decrease in the number of tests ordered (Low: 1.10 vs. $High$: 1.30) and received (0.67 vs. 0.93), as well as a decrease in both congestion cost (1.91 vs. 3.65) and diagnostic cost (9.17 vs. 17.29). The reason is threefold. First, the lower prior makes it more likely that an item passes the first test, which stops the diagnostic process and provides perfect accuracy. Second, while the optimal policy in Low allows for substantial increases in congestion at $k = 1$ and $k = 2$, such increases are unlikely (shaded areas in Figure 1b). Third, even if congestion increases substantially, the DM should make a series of diagnoses without testing at all (at reasonably good odds p_0) to quickly lower congestion to $x = 1$. The extent to which the DM can reap these benefits of the Low uncertainty environment depends on the existence and magnitude of decision bias regarding the three stopping rules (Ⓑefore, Ⓐfter, Ⓦhile) and the cost of such biases in the two environments.

5.2. Results

Table 2 Study 1: Results

		Tests ordered $\bar{O}$			Tests received $\bar{R}$			Wait Cost $\bar{C}$			Diagnostic Cost $\bar{P}$			Total Cost $\bar{C}+\bar{P}$		
Condition	N	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*
High	23	1.30	$>^*$	.96	.93	$>^\dagger$	.82	3.65	$<^*$	8.61	17.29	$<^*$	22.91	20.94	$<^*$	31.52
Low	29	1.10	$>^*$	.65	.67	$>^*$	.55	1.91	$<^*$	3.71	9.17	$<^*$	13.59	11.08	$<^*$	17.29

Notes: $^*p < 0.01$;$^\circ p < 0.05$;$^\dagger p < 0.1$. Two-sided Wilcoxon signed-rank test.

Cost (Hyp. 4). Table 2 presents the aggregate results. Using subject-level averages as the unit of analysis we observe that the total cost is 51% higher than optimal in $High$ (31.52 vs. 20.94), and 56% higher than optimal in Low (17.29 vs. 11.08). The data show that DMs order fewer tests than theoretically optimal and as a result incur higher-than-optimal diagnostic costs, in $High$ ($\bar{P}$: 22.91 vs. 17.29) and Low ($\bar{P}$: 13.59 vs. 9.17). However, they simultaneously incur higher-than-optimal congestion cost, in $High$ ($\bar{C}$: 8.61 vs. 3.65) and Low ($\bar{C}$: 3.71 vs. 1.91). We next study the reasons for this partial lack of support for Hypothesis 4.

Undertesting (Hyp. 1A - Rule Ⓑ). We observe the predicted higher-than-optimal diagnostic costs because of undertesting at low congestion levels where testing is relatively inexpensive. For

example, when the system is not congested at $x=1$, DMs receive on average 0.89 tests in $High$ and 0.33 tests in Low (Figure 4), which is significantly below the respective optimal stopping thresholds $\bar{k}(1)=3$ and $\bar{k}(1)=4$. To further study the mechanisms underlying the observed early stopping behavior at $x=1$, Table 3 displays stopping decisions at different times of the diagnostic process. The data shows that DMs make a large fraction of these diagnoses without testing at all (Rule Ⓑ). In $High$, 15% of diagnoses at $x=1$ are made without testing. The performance implication of such pre-mature stopping is substantial. The expected diagnostic penalty cost of diagnoses without testing is \$40 (and \$60 if the DM mistakenly diagnoses "faulty"), which compares quite unfavorably with the optimal *total* expected cost of \$20.95 from Table 2. Similarly, in Low, 47% of diagnoses at $x=1$ are made without testing. The expected diagnostic penalty cost of such diagnoses is \$14 (and \$86 if the DM mistakenly diagnoses "faulty"), which again compares unfavorably with the optimal *total* expected cost of \$11.01.

Undertesting (Hyp. 1A - Rule Ⓐ). Table 3 documents how pre-mature stopping after receiving a test result contributes further to the overall undertesting pattern at $x=1$. DMs in $High$ are likely to stop after failing one (56%) or two tests (88%), short of the optimal $\bar{k}(1)=3$. Similarly, DMs in Low are likely to stop after failing one (40%), two (70%), or three tests (100%), short of the optimal $\bar{k}(1)=4$. This early stopping after test receipts is particularly hurtful to performance in Low ($p_0=0.14$), where failed tests early in the diagnostic process actually increase diagnostic uncertainty ($p_1=0.4$, $p_2=0.73$).

Table 3 Stopping after receiving test result k at $x=1$.

		p_k				
Condition		0.14	0.4	0.73	0.91	0.98
			$k=0$ …	$k=1$ …	$k=2$ …	$\bar{k}=$ ❸ …
$High$	observations		1,745	305	44	4
	stop (in %)		**15%**	**56%**	**88%**	**100%**
		$k=0$ …	$k=1$ …	$k=2$ …	$k=3$ …	$\bar{k}=$ ❹ …
Low	observations	3,579	251	67	8	
	stop (in %)	**47%**	**40%**	**70%**	**100%**	

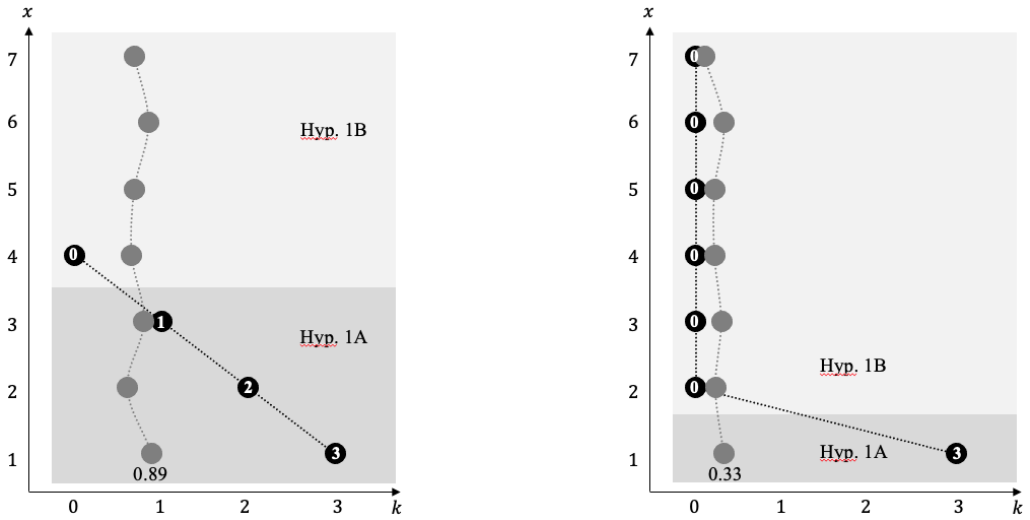
Notes: '…' indicates processes that transition to $x>1$ between test results k and $k+1$.

Undertesting and over-congestion. Because systematic undertesting at low congestion levels in itself would keep system congestion low, the observed higher-than-optimal congestion cost suggests that DMs reach high and theoretically infeasible congestion levels more often, and stay there for longer, than permitted under the optimal policy. This pattern would arise if DMs propensity to stop *before* ordering or *after* receiving a test (Rules Ⓑ and Ⓐ) does not increase at higher congestion x, or if DMs fail to abort as x increases *while* waiting for a test result (Rule Ⓦ). Essentially,

the observed higher-than-optimal congestion cost must be the result of insufficient sensitivity to congestion, which can manifest itself in the violation of Hypothesis 2 (Sensitivity) or Hypothesis 3 (Path Independence), or both.

Congestion Sensitivity (Hyp. 2). We predicted that DMs stop too early (Hypothesis 1A) at low congestion levels that require testing, but are generally sensitive to congestion (Hypothesis 2), for diagnostic processes that do not include passed tests or intermittent arrivals. For such "uninterrupted" (by congestion increases) processes in the data of condition $High$, Figure 4 displays the average number of tests received for diagnoses made at each congestion level x. A visual inspection suggests that stopping behavior is insufficiently sensitive to congestion.

Figure 4 Tests received ($\bar{R}$) for sequences without intermittent arrivals: Predictions (●) and Data (●)



(a) $High$ uncertainty ($p_0 = 0.40$) **(b)** Low uncertainty ($p_0 = 0.14$)

Overtesting (Hyp. 1B). When DMs do not adjust their stopping policy (much) to varying levels of congestion, the immediate implication is that they overtest at congestion levels where they should not test at all. For condition $High$, Figure 4 shows that the average number of tests received is larger for any congestion level $x \geq 4$, where $\bar{k}(x) = 0$. Similarly, in Low, DMs incorrectly start or continue about 30-40% of diagnostic processes when the optimal policy predicts the DM stops the process before receiving any test result, $\underline{k}(x) = 0$. Overall, in support of Hypothesis 1B, the data shows that DMs overtest in states with high congestion levels for which the optimal policy predicts no testing at all ($\underline{k}(x) = 0$ for $x > 1$ in Low, $\bar{k}(x) = 0$ for $x > 3$ in $High$). Because the optimal policy leaves room only for overtesting in such states, the more important question concerns the mechanisms that grow congestion to these levels in the first place, given that most of these states cannot be reached often (Low) or at all ($High$) under the optimal policy (see Figure 1).

Path-Dependence (Hyp. 3 - Rule Ⓦ). A plausible explanation is that DMs are insufficiently sensitive to (increases in) congestion, and fail to abort diagnostic processes as the system is transitioning to higher congestion levels. In violation of Hypothesis 3, DMs may continue the search when they see congestion increase from $(x-1,k)$ to (x,k) while waiting for a test result (Rule Ⓦ), even though they may stop in the same state after receiving a test result (Rule Ⓐ $(x,k-1)\rightarrow(x,k)$). To formally test this idea, we next estimate a probit regression model with the τ's decision by subject i as the dependent variable ($y_{i\tau}$=1 if stop and diagnose, 0 else),

$$\Pr(y_{i\tau}=1)=\Phi(Constant+\alpha_x x_{i\tau}+\alpha_k \mathbf{K}_{i\tau}+\alpha_{xk} x_{i\tau}\mathbf{K}_{i\tau}+\alpha_P Path_{i\tau}+\upsilon_c), \tag{2}$$

which includes a random effect υ_c to account for subject-level heterogeneity in stopping behavior. We let $x_{i\tau}$ denote the congestion level in state $s_{i\tau}=(x_{i\tau},k_{i\tau})$. Because the effect of k on stopping behavior is unlikely to be linear, we use a set of dummy variables $\mathbf{K}_{i\tau}$ that capture the number of (positive or negative) tests received for the item under service. For example, with the *Constant* providing the baseline at $k=0$, $K_1=1$ if an item has failed exactly one test. Together, $x_{i\tau}$ and $\mathbf{K}_{i\tau}$ fully describe the system at subject i's decision epoch τ. The main variable of interest is $Path_{i\tau}$, which is 1 if the system transitioned into state s by arrival while waiting for a test result ($x_\tau=x_{\tau-1}+1$ for $x_{\tau-1}>0$), and 0 else. Hypothesis 3 predicts that $Path_{i\tau}$ has no effect on stopping behavior. We also include an interaction term $x_{i\tau}\mathbf{K}_{i\tau}$, mainly to assess whether $Path_{i\tau}$ has descriptive validity even after controlling for (non-hypothesized) effects such as k moderating a DM's sensitivity to congestion x.

Table 4 presents the results of estimating Model (2) on the data of each condition, including nested versions that exclude $Path_{i\tau}$ or the interaction term. We also carried out the same estimations for the second half ($t=201-400$), to assess whether effects are robust to learning. The variables $\mathbf{K}_{i\tau}$ organize the data sensibly, indicating that the stopping probability is highest when DMs have received a passed test (for which p_k=0), or when two ($High$) or three (Low) successive failed test have moved the posterior to p_k=0.91. The positive estimates for x indicate that decision making is not entirely insensitive to congestion (Hypothesis 2), with the effect growing stronger when we constrain ourselves to the data from the second half ($t=201-400$). Importantly, the negative estimates for $Path_{i\tau}$ show path-dependent behavior in violation of Hypothesis 3. Essentially, DMs are less likely to stop in a given state (x,k) when they are waiting for a test result (Rule Ⓦ), as opposed to when they entered the state after receiving a test result (Rule Ⓐ).

Table 4 Study 1: Estimation Results

DV: y_{is}	$t=1-400$						$t=201-400$					
	Low			*High*			*Low*			*High*		
IV	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
Constant	.06	.20	.15	-1.84^{*}	-1.67^{*}	-1.71^{*}	$-.01$	.04	.07	-2.20^{*}	-1.97^{*}	-1.93^{*}
Path	–	-1.24^{*}	-1.26^{*}	–	$-.66^{\circ}$	$-.64^{\circ}$	–	-1.51^{*}	-1.50^{*}	–	$-.94^{*}$	$-.93^{*}$
$K_{<0}$	2.66^{*}	2.39^{*}	2.84^{*}	4.08^{*}	3.89^{*}	3.12^{*}	2.85^{*}	2.59^{*}	2.51^{*}	4.33^{*}	4.06^{*}	3.32^{*}
K_1	$-.47^{\dagger}$	$-.53^{\circ}$	$-.65^{\circ}$	2.12^{*}	2.01^{*}	2.29^{*}	$-.98^{*}$	-1.05^{*}	-1.53^{*}	2.27^{*}	2.13^{*}	2.05^{*}
K_2	1.29^{*}	1.14^{*}	1.23^{*}	4.07^{*}	3.86^{*}	4.22^{*}	1.09^{*}	$.98^{*}$	1.27^{*}	4.91^{*}	4.60^{*}	5.01^{*}
K_3	4.21^{*}	3.88^{*}	2.90°	–	–	–	4.09^{*}	3.86^{*}	3.46^{*}	–	–	–
x	$-.055$	.002	.02	$.06^{\circ}$	$.07^{*}$	$.08^{\circ}$	$.10^{\circ}$	$.21^{*}$	$.20^{*}$	$.11^{*}$	$.13^{*}$	$.12^{*}$
$x \cdot \mathbf{K}$	*No*	*No*	*Yes*	*No*	*No*	*Yes*	*No*	*No*	*Yes*	*No*	*No*	*Yes*
LPL	-6,190	-5,955	-5,939	-5,137	-5,075	-5,042	-2,831	-2,682	-2,677	-2,561	-2,501	-2,493
Obs.		11,669			11,407			5,756			5,859	

Notes: $^{*}p<0.01;^{\circ}p<0.05;^{\dagger}p<0.1$

5.3. Discussion

The data exhibit qualitatively similar decision biases and performance loss in both conditions. Contrary to Hypothesis 4, we find that DMs accumulate higher-than-optimal cost on both dimensions of the accuracy/congestion trade-off. Our analyses shed light on the mechanisms behind this result, and allow for a cautious attempt at predicting the effect of observed individual-level bias on how the system occasionally cycles between low and high congestion regimes.

At low system congestion, we find that DMs settle for too few tests, in support of Hypothesis 1A. Because undertesting at low congestion levels in itself should keep congestion low, why then do DMs incur substantially higher-than-optimal congestion cost? Our results point to two related mechanisms. First, in support of Hypothesis 2, stopping behavior is insufficiently sensitive to congestion level. The implication is that DMs continue testing at congestion levels where it is optimal to diagnose without a test, at the risk of increasing congestion further. Notably, this includes congestion levels that the system would never reach under the optimal policy. Second, contrary to Hypothesis 3, DMs are less likely to stop the diagnostic process when they see an increase in congestion level while waiting for a test result (Rule Ⓦ) than after they receive a test result (Rule Ⓐ). This violation of the path-independent property of the optimal dynamic decision rule allows congestion to grow to levels that should remain inaccessible at optimality. Once these biases have contributed to taking the system to higher congestion levels, DMs struggle with reducing it. On the one hand, path-dependent behavior creates a tendency for further congestion increases. On the other hand, once the system has reached congestion levels not predicted by theory, DMs do not make enough use of rule Ⓑ (Stop before testing) as the most effective lever to reduce congestion. Ironically, the same behavior that tends to keep the system at low congestion levels may prevent

the system from going back to low congestion levels. As a result, system performance may suffer on both dimensions of the accuracy/congestion trade-off. We next study the underlying behavioral reasons and what can be done to improve performance.

6. Study 2: Providing statistical support.

The high diagnostic costs observed in Study 1 result because DMs settle for too few tests at low congestion levels. We predicted this pattern in part because excessive belief adjustments in the direction of a test result may prompt DMs to stop too early with too little diagnostic information. The data from Study 1 lends support for this behavioral driver behind Rule Ⓐ (Stop after testing). Because we design the two experimental conditions such that $p_k^{Low} = p_{k-1}^{High}$, we can directly compare stopping behavior at a particular probability p_k. For example, Table 3 shows that DMs are conditionally more likely to stop after the first test in condition *Low* ($p_1 = 0.4$) than to stop without test in condition *High* ($p_0 = 0.4$), even though diagnostic information is the same in both cases. Arguably, DMs overweigh the signal and quickly stop on the too few test results, and it stands to reason that they perform poorly simply because they cannot correctly calculate the posterior based on the tests they have received.

Study 2 provides a direct test of this conjecture. We implement a treatment that is identical to Study 1, with the only difference that we show DMs the correct posterior p_k throughout the task, readily displayed in the diagnostic decision buttons "Faulty. Probability: $[p_k\%]$" and "Good: Probability: $[1 - p_k\%]$", respectively (see Appendix EC.1.4). We implement the new treatment *Show* for both *Low* and *High* uncertainty conditions, and we use the label *Hide* to refer to the Study 1 treatments where DMs were not able to see the correct posteriors.

6.1. Results

Performance. Table 5 presents the results from Study 2 (*Show*) and includes the results from Study 1 (*Hide*) for easy comparison. Our main observation is that performance does not improve when DMs have access to the correct posterior, with no significant performance differences between *Hide* and *Show* (Mann-Whitney-U tests for $\bar{C}$, $\bar{P}$, $\bar{C} + \bar{P}$). To study the reasons for this no-effect, note that overall performance depends on the diagnostic decisions conditional on stopping the process (Proposition 1a) as well as the stopping decisions themselves (Proposition 1b).

Table 5 The impact of statistical support on performance

			Tests ordered $\bar{O}$			Tests received $\bar{R}$			Wait Cost $\bar{C}$			Diagnostic Cost $\bar{P}$			Total Cost $\bar{C}+\bar{P}$		
Cond.	Treat.	N	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*
High	*Show*	26	1.30	$>^*$	.95	.93	$>^\circ$	.81	3.65	$<^*$	12.90	17.29	$<^*$	21.19	20.94	$<^*$	34.09
High	*Hide*	23	1.30	$>^*$	.96	.93	$>^\dagger$	.82	3.65	$<^*$	8.61	17.29	$<^*$	22.91	20.94	$<^*$	31.52
Low	*Show*	23	1.11	$>^*$	.64	.67	$>^*$	.52	1.91	$<^\dagger$	3.67	9.09	$<^*$	12.18	11.01	$<^*$	15.86
Low	*Hide*	29	1.10	$>^*$	.65	.67	$>^*$	.55	1.91	$<^*$	3.71	9.17	$<^*$	13.59	11.08	$<^*$	17.29

Notes: $^*p<0.01$;$^\circ p<0.05$;$^\dagger p<0.1$. Two-sided Wilcoxon signed-rank test. *Hide* treatments from Study 1.

Diagnostic Decisions. For each diagnosis $d_{i\tau}$ made by subject i with decision τ, we count a diagnosis as correct ($d^*_{i\tau}=1$) if it coincides with the optimal diagnosis according to Proposition 1a, and as incorrect otherwise ($d^*_{i\tau}=0$). We calculate the fraction of correct diagnosis *conditional on* stopping at p_k. For the $High$ uncertainty environment, diagnostic accuracy appears to be marginally higher in *Show* than in *Hide*, at $p_1=0.73$ (0.94 vs. 0.92), and $p_2=0.91$ (1 vs. 0.95). For the *Low* uncertainty environment, diagnostic accuracy is higher in *Show* than in *Hide* at $p_1=0.40$ (0.42 vs. 0.24), but not at $p_2=0.73$ (1 vs. 1). To formally test these descriptive observations, we estimate for each condition (*Low*, $High$) a probit regression model,

$$\Pr(d^*_{i\tau}=1)=\Phi(\alpha_0+\alpha_1\mathbf{K}_{i\tau}+\alpha_2 Show_i+\alpha_3 t+\upsilon_c), \tag{3}$$

where the set of dummy variables $\mathbf{K}_{i\tau}$ controls for the posterior p_k at which the diagnosis is made, and t captures possible learning effects. The main estimate of interest is for the variable $Show_i$. The results show that access to the statistical posterior has no material impact on the conditional (on p_k at the point of stopping) quality of diagnostic decisions, on average, in *Low* ($\alpha_2=.01, p=.91$) and in $High$ ($\alpha_2=.20, p=.16$). In light of these results, it is not surprising that DMs do not overall incur lower diagnostic cost $\bar{P}$ when they have direct access to the correct posterior in treatments *Show* than when left to their own cognitive devices in treatments *Hide*.

Stopping Decisions. A cursory investigation of stopping decisions in different system states shows no systematic differences between the $Hide$ and $Show$ treatments, with one notable exception for *Low* uncertainty environments: when initial test results have effectively pushed the diagnostic process into system states with *higher* diagnostic uncertainty (compared to $p_0=0.14$), DMs are less likely to mistakenly stop the diagnostic process in $Show$. For example, looking only at stopping decision when congestion is $x=1$ (Table EC.3 in Appendix B), we observe that DMs in the *Low* uncertainty condition are less likely to mistakenly stop when they see the correct posterior ($Show$) than when they do not ($Hide$), both at $p_1=0.40$ (21% vs. 40%) and at $p_2=0.73$ (54% vs. 70%).

6.2. Discussion.

We observe that access to the correct statistical posterior may help DMs when initial diagnostic information has pushed the system into states where diagnostic uncertainty is high and judgment bias is likely to lead the DM to mistakenly stop and (conditional on stopping) make incorrect diagnoses. However, the benefit does not show up in the aggregate results, where we register a performance loss of 44% in *Low* (total cost: 15.86 vs. 11.01) and 63% in *High* (34.09 vs. 20.94). The likely reason is that DMs rarely enter system states where access to the correct statistical posterior can help them make better diagnostic decisions. Overall, Study 2 shows that providing access to the correct posterior does not help DMs manage congestion better in the environments we implemented. Study 2 (*Show*) further corroborates the key results from Study 1 (*Hide*). We observe again that DMs incur both higher-than-optimal diagnostic cost and higher-than-optimal congestion cost (Table 5). And we again observe path-dependent behavior under which DMs are more likely to stop after receiving a test result than after experiencing an increase in system congestion (see Table EC.2 in Appendix B for details).

7. Study 3: Managing Diagnostic Processes without Congestion.

Study 3 sheds light on how the mere presence of congestion may (or may not) change the way DMs approach the fundamental sequential search and hypothesis testing problem underlying the diagnostic task. A striking result from Studies 1 and 2 is that DMs severely undertest even when testing is cheap at low congestion levels such as $x = 1$. Our results thus far show that DMs undertest because they may like the odds at p_0 (Rule Ⓑ), or because they may excessively update p_0 based on few test results (Rule Ⓐ). That is, they may undertest for reasons that are not directly linked to the cost of future congestion. As a complementary reason, we posit that DMs may overestimate the future congestion-related cost of testing. Testing this conjecture would require either the manipulation of the decision biases that drive costs as system congestion grows, or to manipulate these costs directly. Because the controlled elimination of decision bias at $x > 1$ is impractical, we instead decrease the expected congestion-related cost of testing at $x = 1$ directly.

7.1. Design and Implementation.

For the *Low* and *High* uncertainty conditions from Study 2 (henceforth, *Congestion*), we implement a *No Congestion* treatment that removes congestion entirely. DMs in *No Congestion* never see a queue build up during the diagnostic process and only incur costs for the current item. Specifically, we run the same algorithm and display the same interface in all our studies, with the only differences that in Study 3, new arrivals are not displayed and their impact on cost not accounted for. In this study, DMs are only told that an ordered test result returns with probability 1/2 at each time

epoch. As a result, DMs experience the exact same sequence of time epochs, and are presented the same choices at each time epoch in all our studies, whether congestion is present or not.

As a result, DMs incur search cost $c = \$1$ each time they decide to either order a new test, or continue to wait for a test that they had ordered, such that the total search cost until time epoch t reduces to $C_t = ct$. The expected number of time epochs for a test's result to come back is equal to $1/(1-\lambda)$, which yields the expected cost per test $c/(1-\lambda) = \$2$ for our parameter values.

Optimal stopping in the absence of congestion follows a simple threshold policy, under which the DM should run tests as long as the tests fail and the number of failed tests is less or equal to threshold $\bar{k}$ (Section 3). Because the system without congestion is equivalent to a congested system at $x = 1$ with no new arrivals, we would not predict any between-treatment differences in stopping behavior under the optimal policy. Importantly, we would also not predict any between-treatment differences based on any judgment and decision-making biases that are not inherently linked to future congestion, such as the excessive updating of p_k.

7.2. Results

Performance. We had predicted that undertesting would lead to lower-than-optimal search cost at the expense of higher-than-optimal diagnostic cost in settings *with* congestion (Hypothesis 4), in parts because of results from related search settings *without* congestion. Although a rational DM cannot suffer on both dimensions of the accuracy/search cost trade-off in search settings *without* congestion, the results from Studies 1 and 2 suggest it is possible that a boundedly rational DM might. The *No congestion* treatment offers a direct test of this (Table 6). For both uncertainty conditions, DMs tend to order ($\bar{O}$) and receive ($\bar{R}$) fewer tests than optimal, in support of Hypothesis 1A. This undertesting pattern translates into higher-than-optimal diagnostic cost $\bar{P}$ and lower-than-optimal search cost $\bar{C}$, which corroborates the early stopping results from the empirical literature on search problems without congestion.

Table 6 The impact of congestion on performance

			Tests ordered $\bar{O}$			Tests received $\bar{R}$			Wait Cost $\bar{C}$			Diagnostic Cost $\bar{P}$			Total Cost $\bar{C}+\bar{P}$		
Cond.	Treatment	N	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*	*Pred.*		*Obs.*
High	*No congestion*	13	1.95	$>$	1.78	1.95	$>^{\dagger}$	1.72	3.83	$>^{\dagger}$	3.38	.97	$<^{*}$	6.20	4.80	$<^{*}$	9.57
High	*Congestion*	26	1.30	$>^{*}$	.95	.93	$>^{\circ}$	.81	3.65	$<^{*}$	12.90	17.29	$<^{*}$	21.19	20.94	$<^{*}$	34.09
Low	*No congestion*	25	1.62	$>^{*}$	1.14	1.62	$>^{*}$	1.06	3.19	$>^{*}$	2.07	0	$<^{*}$	6.60	3.19	$<^{*}$	8.68
Low	*Congestion*	23	1.11	$>^{*}$	.64	.67	$>^{*}$	.52	1.91	$<^{\dagger}$	3.67	9.09	$<^{*}$	12.18	11.01	$<^{*}$	15.86

Notes: $^{*}p < 0.01$;$^{\circ}p < 0.05$;$^{\dagger}p < 0.1$. Two-sided Wilcoxon signed-rank test. *Congestion* treatments from Study 2. For $\bar{O}$ and $\bar{R}$, all differences between *Congestion* and *No Congestion* are significant (Mann-Whitney-U, $p < 0.01$).

Stopping rules. Table 7 illustrates how the overall undertesting pattern arises. First, DMs stop the diagnostic process before testing at all (Rule Ⓑ), and more so at the relatively better diagnostic

prior odds of the *Low* environment. Second, DMs stop after having received too little test information (Rule Ⓐ). The data lends further support to the idea that DMs overreact to, and overweigh, small samples of test results. For example, DMs are more likely to stop based on a single test in *Low* than without any test in *High* (22% vs. 1%), even though $p_k = 0.4$ in both cases. In fact, Table 7 suggests that the number of test results itself (k) is a better predictor of stopping than the probability p_k that the item is faulty, despite the fact that DMs have access to p_k throughout the experiment. Third, DMs sometimes prematurely stop while waiting for a test to return (Rule Ⓦ), which is *never* optimal in *No congestion* settings, and necessary only at $x > 1$ in *Congestion* settings where waiting for a test implies an increase in congestion.

Table 7 Stopping after test result k and while waiting (...) at $x = 1$

			p_k											
			0.14		0.4		0.73		0.91		0.98		0.99	
Cond.	Treatment				$k=0$	...	$k=1$	...	$k=2$	...	$k=$ ❸	...	$k=4$	...
High	*No congestion*	observations			1,525	1,391	750	641	526	320	294	55	51	3
		stop (in %)			**1%**	4%	**12%**	2%	**30%**	15%	**85%**	39%	**97%**	33%
High	*Congestion*	observations			2,019		330		39		1		-	
		stop (in %)			**20%**		**65%**		**82%**		**100%**		-	
			$k=0$	...	$k=1$	...	$k=2$	...	$k=3$	...	$k=$ ❹	...	$k=5$	...
Low	*No congestion*	observations	3,999	2,583	862	706	454	324	231	49	53	9	-	-
		stop (in %)	**21%**	20%	**22%**	6%	**27%**	2%	**75%**	38%	**86%**	50%		
Low	*Congestion*	observations	2,988		218		57		11		-		-	
		stop (in %)	**48%**		**21%**		**54%**		**100%**		-		-	

Notes: *Congestion* treatments from Study 2. In *Congestion*, waiting (...) implies an increase in congestion ($x > 1$)

(Anticipated) congestion matters. In the absence of congestion and related congestion cost, Table 6 shows that DMs test more (*Obs.* $\bar{O}$ and $\bar{R}$), as they should (*Pred.* $\bar{O}$ and $\bar{R}$). To draw a sharper contrast, we leverage the fact that theory predicts the same stopping behavior for *No congestion* as it does for a *Congestion*-prone environment that is at $x = 1$. The data is at odds with this prediction (Table 7) - DMs are more likely to stop in the presence of congestion than in its absence, at almost any level of k. The implication is that DMs in *Congestion*-prone environments undertest at low congestion levels (here, $x = 1$) in part because of the same judgment and decision biases one can expect in *No congestion*, and in part because of the anticipation of congestion-related cost caused by their own inability to manage the system well at higher congestion level.

8. Study 4: Distinguishing & debiasing performance loss drivers

To further our understanding of the reasons for the substantial performance loss observed in our studies, and towards the design of debiasing mechanisms, we note that sub-optimal performance has roots in two broad classes of bias.

Poor strategy. Our results thus far show that stopping thresholds are *on average* insufficiently sensitive to congestion (illustrated in Figure 4), with the implication that DMs do not test enough, while still allowing the system to reach congestion levels that are infeasible under the optimal policy. In other words, performance may suffer from poor choice of stopping thresholds.

Poor execution. Performance may also suffer when DMs fail to make consistent decisions, i.e., when they do not always make the same decision in a given system state (x, k). One example for such variability in decision making is the systematic path-dependency that we observe in our data. Note that even a poor strategy, if executed consistently, would adhere to the normatively compelling principle of path-independence in decisions. Variability in choices can also result from randomness in the decision making process. Indeed, the prediction of Hypothesis 1A (Undertesting at low congestion) builds on prior evidence that random choice can account for costly early-stopping bias observed in search settings without congestion (Bearden and Murphy 2007), and we made a similar argument with regard to Hypothesis 1B (Overtesting at high congestion levels).

Understanding whether performance loss is due to poor strategy or due to poor execution is important because each category of bias is likely to require different debiasing techniques (Arkes 1991). Since it is difficult to distinguish poor strategy from poor execution based on the data from Studies 1 and 2 alone (without relying on econometric exercises and the assumptions that typically go along), Study 4 makes an attempt to answer the question experimentally.

8.1. Design and Implementation

To address the question of poor strategy, we directly elicit the DM's stopping thresholds, rather than estimating them from our choice data (e.g., using Equation 2). To address the question of poor execution, we constrain the DM's tendency to exhibit variability in the execution of their strategy.

We implement three treatments in the *High* uncertainty condition with the statistical posteriors being shown throughout (*Show* treatments from Studies 2 and 3). Each treatment has three parts (Table 8). Part A is identical for all three treatments, and DMs perform the diagnostic task from Studies 1 and 2 for $T = 100$ time epochs. Parts B and C implement our key interventions.

Table 8 Study 4 Design

Treatment	N	Part A: $t = 1 - 100$	Part B	Part C: $t = 1 - 300$
Base	59	Diagnostic task	Neutral filler task	Diagnostic task
Recommend	59	Diagnostic task	Strategy elicitation	Diagnostic task with own strategy as recommendation
Commit	61	Diagnostic task	Strategy elicitation	Diagnostic task with own strategy as commitment

Strategy elicitation (Part B). In treatments *Recommend* and *Commit*, we elicit from subjects their diagnostic testing strategy, defined as a set of stopping thresholds $S_i \equiv \{\bar{k}_i(1), \bar{k}_i(2), ...\}$. In short, with more procedural details relegated to Appendix EC.1.5, DMs were first prompted to indicate the maximum number of items $\bar{x}$ that they would allow in the system. Next, at each congestion level from $x = 1$ to $\bar{x} - 1$, subjects indicated the number of failed tests $\bar{k}(x)$ after which they would stop the process and diagnose the item. The strategy elicitation in part B provides for each subject i a strategy $S_i \equiv \{\bar{k}_i(1)...\bar{k}_i(\bar{x}_i)\}$. To replicate the cognitive efforts that the other two treatments impose on the DMs in part B, treatment *Base* implements a simple but tedious filler task that requires DMs to count the number of items on the screen.

Strategy implementation (Part C). DMs perform the same diagnostic task from part A, for $T = 300$ time epochs. In treatment *Base*, the system ran exactly as in Studies 1 & 2, without any intervention. In treatment *Recommend*, the system displayed the DM's own strategy $\bar{k}_i(x)$ as a recommendation for the current state (x, k), but DMs were allowed to deviate from the recommendation. In treatment *Commit*, the system would implement the strategy from part B, without the option to deviate from it. In order to keep the overall time spent similar across treatments, DMs in *Commit* still have to "click through" part C.

Distinguishing and debiasing loss drivers. A comparison between *Commit*/*Recommend* and *Base* allows us to assess the impact of providing DMs with the incentive and time to carefully think through a strategy after having gained some task experience in part A. Furthermore, our treatments are designed to distinguish performance loss due to poor strategy from performance loss due to poor execution. Treatment *Commit* allows assessing the magnitude of performance loss due to poor strategy (biased $\bar{k}_i \neq \bar{k}$). Because *Commit* eliminates poor strategy execution by design, any performance loss would be entirely due to poor stopping strategies defined in part B. In contrast, treatment *Recommend* allows assessing the magnitude of performance loss due to poor strategy execution, by comparing the observed performance of *Recommend* with the counterfactual performance had DMs perfectly executed their own strategy S_i from part B. Indeed, DMs may perform worse than their own strategy S_i when they subsequently fail to apply their stopping thresholds consistently, or when they make incorrect diagnostic decisions conditional on having stopped the diagnostic process. Treatment *Recommend* further provides a window into the difficulties DMs may experience when trying to execute a diagnostic testing strategy consistently, in an environment that is everything but (consistent).

Implementation. We recruited participants from the regular subject pool associated with the experimental laboratory at the University of Hamburg, Germany. The experiment was implemented in SoPHIE (Hendriks 2012) and conducted online (see Appendix EC.1.4 for a screenshot of the user interface). Upon arrival, subjects entered a virtual wait room and were allocated randomly to one

of the three treatments. As in Studies 1-3, subjects were compensated based on their performance over $T = 400$ time epochs, measured by the average cost per diagnosed item in parts A and C.

8.2. Results

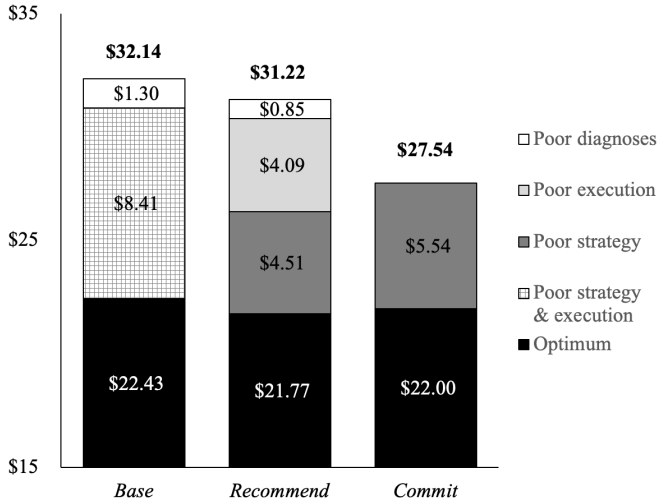
Figure 5 displays the treatment level averages of per-diagnosis total cost. We include the optimal cost as a benchmark, derived from applying the optimal policy (Figure 1a) to the same sample paths of random events faced by the DMs. Note that the distribution of sample paths is not perfectly balanced across the three treatments, and the optimal benchmarks differ slightly as a result. To control for this, we define $LossC_i = (\bar{C}_i + \bar{P}_i) - (\bar{C}(\bar{S}, \sigma_i) + \bar{P}(\bar{S}, \sigma_i))$ as the difference between DM i's observed total cost in part C and the counterfactual total cost from applying the optimal policy $\bar{S}$ on the same sample path σ_i played by the DM. To formally test the following observations, we then estimate a simple regression model,

$$LossC_i = \alpha_0 + \alpha_1 Recommend_i + \alpha_2 Commit_i + \alpha_3 \mathbf{\Sigma} + \alpha_4 LossA_i, \tag{4}$$

which uses treatment *Base* as the baseline, and includes a set of dummy variables $\mathbf{\Sigma}$ to control for sample path σ_i. We also include $LossA_i$ (defined similar to $LossC_i$) to control for a DM's performance in part A.

Does strategy elicitation help? We find that total cost in *Commit* is significantly lower than in *Base* (27.54 vs. 32.14, $p < .01$). While total cost in *Base* is about 43% higher than optimal (32.14 vs. 22.43, $p < .01$), it is about 25% higher than optimal (27.54 vs. 22.00, $p < .01$) in *Commit*. This captures the value of defining a strategy *and* consistently applying it. However, total cost in *Recommend* is larger than in *Commit* (31.22 vs. 27.54, $p < .01$) and not much lower than in *Base* (31.22 vs. 32.14, $p = 0.69$). Given our experimental design, the reason for this non-effect of *Recommend* must either be poor strategy definition or poor strategy execution, or both.

Figure 5 Part C ($t = 1 - 300$): Per-diagnosis total cost $\bar{C} + \bar{P}$ and performance loss decomposition



Strategy definition. It is possible that DMs define better strategies in part B of *Commit* than in *Recommend* for a simple incentive reason - while DMs in *Commit* know that their strategy from part B will directly affect their earnings from part C, DMs in *Recommend* know that they will be able to freely overwrite their own strategy. To directly test this conjecture, and tease apart performance loss from a poor strategy and performance loss from *not* implementing the (poor) strategy consistently, we next calculate the counterfactual performance of the strategies we elicited in part B. Specifically, we apply each subject i's strategy S_i from part B to the sample path σ_i. This yields for treatment *Recommend* the counterfactual total cost $\bar{C}_i(S_i,\sigma_i)+\bar{P}_i(S_i,\sigma_i)$. For *Commit*, by construction of the treatment, the counterfactual performance of S_i is the actual performance of subject i in part C. Based on these calculations, we can precisely measure the loss from poor strategy as $LossC(S_i) = (\bar{C}_i(S_i,\sigma_i) + \bar{P}_i(S_i,\sigma_i)) - (\bar{C}(\bar{S},\sigma_i) + \bar{P}(\bar{S},\sigma_i))$.

We find that the counterfactual total cost of DMs' strategies S_i is substantially lower in *Recommend* than in *Base* (4.51+21.77=26.28 vs. 32.14, $p<.01$), and also lower in *Commit* than in *Base* (27.54 vs. 32.14, $p<.01$). However, the counterfactual total cost are higher than optimal in both cases (*Recommend*: 26.28 vs. 21.77, $p<.01$; *Commit*: 27.54 vs. 22.00, $p<.01$). Importantly, the data shows no difference in counterfactual costs between *Recommend* and *Commit* (26.28 vs. 27.54, $p=.45$). We could probably remove the rest of this paragraph. For robustness, we repeat this exercise and calculate the counterfactual performance of each elicited strategy S_i from treatments *Recommend* and *Commit*, by applying each S_i to all sample paths used in our experiment. Specifically, we apply each subject i's strategy S_i to each of the $\sigma = 1..\Sigma$ sample paths of pre-generated stochastic outcomes used in part C of the experiment. This yields counterfactual performance measures for congestion cost $\bar{C}_s(S_i)$, penalty cost $\bar{P}_s(S_i)$ and total cost $\bar{C}_s(S_i)+\bar{P}_s(S_i)$. We then calculate the average performance across all sample paths, e.g., $\bar{C}(S_i) = \frac{1}{\Sigma}\sum_{\sigma=1}^{\Sigma}\bar{C}_\sigma(S_i)$. Based on these subject-level means, we observe that treatment-level total cost is substantially higher than optimal ($\bar{C}^* + \bar{P}^* = 22.25$) in *Recommend* (mean: 26.63, median: 24.04) and in *Commit* (mean: 27.62, median: 25.27). Importantly, the data shows no performance difference between *Recommend* and *Commit*.

Strategy execution - Diagnostic Decisions. Because the strategies S_i in *Recommend* are at least as good as those defined in *Commit*, the observed poor overall performance in *Recommend* implies that DMs do not execute their strategies without mistakes. One possible source of bias are incorrect diagnoses after the decision maker has stopped (correctly or not) the diagnostic process. Indeed, while the design of *Commit* eliminates judgment errors such as a "faulty" diagnosis without a test (at $p_0 = 40\%$) or a "good" diagnosis after a failed first test (at $p_1 = 73\%$), these diagnostic mistakes contribute 1.30 to the performance loss in *Base* and 0.85 in *Recommend*.

Strategy execution - Stopping Decisions. Performance may also suffer from stopping decisions that deviate, systematically or randomly, from the DM's strategy S_i. Although performance-improving deviations are generally possible if the DM had defined a poor strategy to begin with, we had predicted a performance loss because of variability in decision making. The data from *Recommend* allows us to quantify this source of performance loss, by comparing the observed cost (after rectifying incorrect diagnoses in the data) with the counterfactual cost that would result from a flawless execution of the DM's strategy S_i. Figure 5 shows that the cost of inconsistent stopping decisions (Poor execution: 4.09) makes up 43% of the total performance loss of 9.45 (31.22 vs. 21.77), which is nearly as high as the cost of defining suboptimal stopping thresholds (Poor strategy: 4.51). Finally, the cost of defining suboptimal stopping thresholds is not different between *Recommend* and *Commit* (4.51 vs. 5.54, $p = 0.34$).

Task experience and learning. To assess whether DMs learn to improve as they gain experience with the task, we can simply compare performance in part A ($t = 1 - 100$) and part C ($t = 1 - 100$) of treatment *Base* which has no intervention in part B. We find that DMs do not make better decision as they gain experience with the task (total cost: 29.70 vs. 28.42, $p = .34$), but they do appear to change how they solve the fundamental trade off, with congestion cost increasing (5.24 vs. 7.58, $p < .01$) and diagnostic penalty cost decreasing (24.46 vs. 20.84, $p < .01$) in part C.

8.3. Discussion

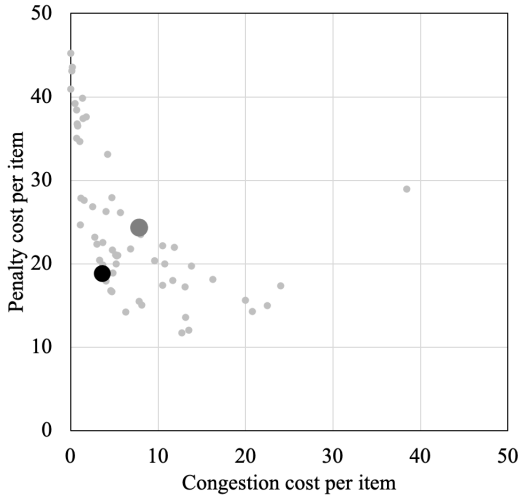
DMs incur significant losses compared to the optimal policy in Study 4 (yielding a 43% increase in total costs in *Base*), which is in line with the findings of Studies 1 and 2. The counterfactual approach we implement in Study 4 reveals that this performance loss is equally due to a poor choice of strategy (i.e. of the stopping thresholds) and the misapplication of this strategy. Specifically, in *Recommend*, poor strategy and poor execution account for about 43% and 48% of the loss, respectively, while the remaining 10% are due to errors in the diagnoses. These findings also highlight the difficulty of debiasing DMs in our context. In particular, *Recommend* does not significantly improve performance compared to *Base*, which indicates that simply making people think about their strategy does not help. One reason for this is that DMs are not consistent with their own strategy as the difference in total costs between *Recommend* and *Commit* reveals. This suggests that forcing DMs to implement their own strategies, even poor ones, can improve performance significantly (by about 14% in our study).

9. Discussion and Conclusion

We study judgment and decision making in diagnostic systems that are prone to costly congestion. Our main finding is that DMs incur lower-than-optimal search cost at the expense of lower-than-optimal diagnostic accuracy in the absence of congestion, but accumulate higher-than-optimal cost

on both dimensions of the accuracy/congestion trade-off in the presence of congestion. Notably, this is not merely an artefact of aggregating across DMs who either keep congestion cost low at the expense of diagnostic cost, or keep diagnostic cost low at the expense of congestion cost. Indeed, about 50% of DMs in our studies underperform on both accuracy and congestion costs (Figure 6 illustrates this for the *Base* treatment of Study 4). These effects are robust to environments with *High* and *Low* levels of uncertainty, regardless of DMs' access to statistical support.

Figure 6 Performance: ● Theory - Treatment avg., ● Data - Treatment avg., • Data - Individual



9.1. Decision Bias: Task Accumulation and Mistake Accumulation

Our results shed light on how this double-sided performance loss relates to individual level bias. We observe substantial *undertesting* at low congestion levels where, due to the relatively small cost of testing, optimal policy dictates the DM tests until she reaches high levels of diagnostic certainty. Although generally consistent with early stopping results from the literature on search without congestion, the underlying mechanisms differ.

At low congestion levels, our data reveals that DMs make too many diagnoses without testing, or on the basis of a very small sample of test results. As a result, we observe higher-than-optimal diagnostic cost for diagnostic processes that remain at low congestion levels. Sacrificing diagnostic accuracy helps lower the cost of diagnostic testing in the absence of congestion (Study 3). But our results show that this trade-off does not translate to diagnostic processes under congestion (Studies 1, 2 and 4). Instead, we observe an over-/undertesting pattern, relative to the theoretical benchmark that would warrant stopping on less diagnostic information as congestion increases. Essentially, DMs are insufficiently sensitive towards congestion.

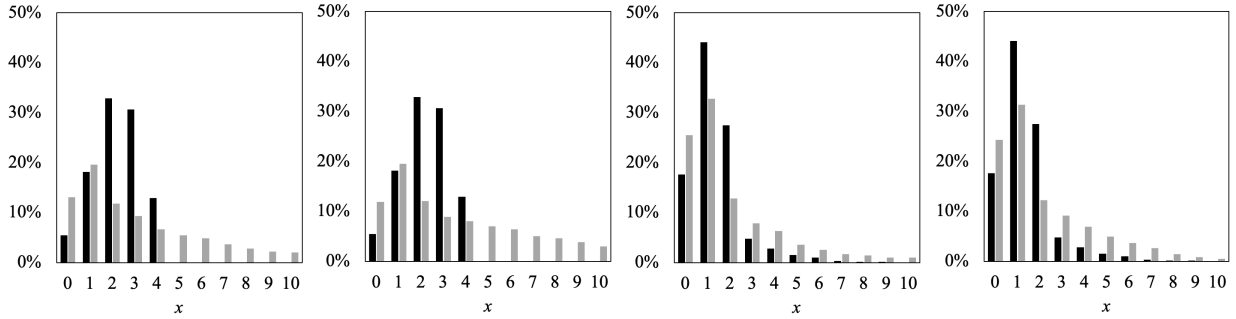
The observed (mis)behaviors are costly, but psychologically sensible, given that the task environment poses taxing demands on optimal decision making. On the one hand, one key structural

property of optimal decision making is intuitive: as the cost of testing increases in congestion, the DM should do less of it, on average. On the other hand, the optimal policy is complicated and includes rules that may run counter to human intuition and preferences. Optimal decision making in our setting includes the assessment of how current decisions affect future decisions and congestion levels, and hence requires the kind of consequential reasoning that DMs are notoriously bad at (Shafir and Tversky 1992). Because future congestion is hard to evaluate, and hence easy to pay insufficient attention to, decision myopia is likely to have a detrimental impact on future performance. Further, the optimal policy requires the abortion of ongoing test procedures as congestion grows (Rule Ⓦ), which may run against the DM's inclination to complete a task that has already started. And once such "task completion bias" has allowed the system to reach excessive congestion levels, reducing congestion requires the DM to diagnose without testing (Rule Ⓑ) more than they might be willing to. It is in this sense that diagnostic systems under dynamic task accumulation are unforgiving to (accumulating) mistakes, much unlike single-shot environments such as the classic newsvendor (Schweitzer and Cachon 2000, Kremer et al. 2010).

9.2. Implications for System Behavior and Performance.

The effect of these biases on system behavior and performance is substantial, resulting in total costs that are between 44% to 63% higher than optimal (Tables 2 and 5). Our data also exhibits more variability in congestion than theoretically predicted (Figure 7).

Figure 7 Variability in system congestion: Theory (■) vs. Data (■) from Studies 1 and 2.



(a) *High* ($p_0 = 0.4$) *Show* **(b)** *High* ($p_0 = 0.4$) *Hide* **(c)** *Low* ($p_0 = 0.14$) *Show* **(d)** *Low* ($p_0 = 0.14$) *Hide*

While spending a larger-than-predicted amount of time at very low congestion levels, DMs at the same time spend a larger-than-predicted amount of time at congestion levels that the system would (almost) never reach under optimal decision making. Related to the high variability in congestion levels, the data shows that the system is idle significantly more often than predicted. In particular, the system is 12-13% idle (vs. predicted 5%) in the *High* conditions, and 24-26% idle (vs. predicted 18%) in the *Low* conditions. When extrapolated beyond the scope of our model and experiments,

which neither explicitly penalize or reward the DM for being idle, the observed underutilization may well warrant management intervention.

9.3. Managerial Implications.

What can management do to debias the behavior we observe, and improve system performance? Our results suggest directions for debiasing mechanisms. In particular, we systematically vary DMs access to the correct posteriors on the item in service, and find that the provision of such statistical information does not significantly affect overall system performance. We also find that providing DMs with incentives and time to carefully think about their strategy does not improve performance either, unless one provides a mechanism that helps DMs stick to their own strategy. In view of the emergence of statistical AI-based systems to help or replace human diagnostic judgments, these findings further suggest the need to debias stopping decisions rather than the diagnostic decisions.

9.4. Contribution and methodological notes.

Our study contributes to the broader discussion about load-dependent server behavior, by studying such behavior in a setting that combines the strengths of experimental, empirical, and theoretical studies. While the standard arguments regarding external validity apply, the controlled design of our study has three main advantages regarding behavioral mechanisms, their qualification as bias, and their links to system-level performance.

First, experimental control helps tease apart reasons for why DMs may or may not engage in task reduction. For example, Batt and Terwiesch (2016) provide anecdotal evidence that doctors may not cut corners (i.e., order fewer tests) when the system is busy, out of ritual, or for "covering the bases". We find the same result, *after* removing rituals and incentives to cover the bases.

Second, building our experiments tightly around the structure and predictions of a formal decision making model, we can qualify whether load-dependency is too much, too little, or just right. In the setting of our study, not only is "task reduction [..] an operational lever that doctors and managers should at least consider" (Batt and Terwiesch 2016), but we can control precisely how much they should consider it. Our results show that DMs are insufficiently sensitive to congestion, and do not cut corners enough as congestion increases.

Third, while generally taking advantage of experimental control, our study embraces (rather than control away) the complexity of the systems that we want to study behavior in. While this complexity introduces empirical challenges reminiscent of field settings, our approach allows us to directly observe system behavior as opposed to making intricate extrapolations from individual decisions observed in isolated system snapshots. In particular, we observe directly how decisions

depend on the dynamics that characterize queuing systems. [4] Indeed, a key contribution of our study is a test of path-dependency in stopping behavior. DMs in our experiment systematically violate path-independence, which is a fundamental and compelling property for decision making in dynamic systems. This could inform future behavioral modeling work as well as econometric specifications of field studies.

Acknowledgments

This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - VE 897/4-1.

[4] To illustrate the point, imagine a simple alternative (to our) experiment to test congestion-dependent behavior with two treatments: *Low* with only 1 item in the system, and *High* with 9 items. Imagine that DMs in such a design stop diagnostic processes after 4 (2) tests in *Low* (*High*). Although such data aligns with the idea that testing decreases in congestion, it is unclear what we can infer about the behavior of dynamic systems. Average congestion is unlikely $\frac{1+9}{2} = 5$, and the average number of tests is unlikely $\frac{4+2}{2} = 3$, unless we are willing to assume the system alternates between the two states evenly. In fact, the experimental design may miss (a) that decisions at *High* congestion levels may not matter (much) if the system (almost) never reaches such levels because of decisions at *Low* congestion levels, and (b) that diagnostic behavior at *High* congestion levels may change depending on how the system arrived at that level.

References

Alizamir S, de Vericourt F, Sun P (2013) Diagnostic accuracy under congestion. Management Science 59(1):157–171.

Alizamir S, de Véricourt F, Sun P (2019) Search under accumulated pressure. Operations Research Forthcoming.

Allon G, Kremer M (2019) Behavioral foundations of queuing systems. Karen Donohue EK, Leider S, eds., The Handbook of Behavioral Operations, 325–366, Handbooks in Operations Research and Management Science (Hoboken, NJ: John Wiley Sons, Inc.).

Arkes H (1991) Cost and benefits of judgment errors: Implications for debiasing. Psychological Bulletin 110(3):486–498.

Batt R, Terwiesch C (2016) Early task initiation and other load-adaptive mechanisms in the emergency department. Management Science 63(11):3531–3551.

Bearden J, Murphy R (2007) On generalized secretary problems. Abdellaoui M, Luce R, Machina M, Munier B, eds., Uncertainty and Risk: Mental, Formal and Experimental Representations (New York: Springer).

Bertsekas DP (2005a) Dynamic programming and optimal control, volume 2 (Athena scientific Belmont, MA), 3 edition.

Bertsekas DP (2005b) Dynamic programming and optimal control, volume 1 (Athena scientific Belmont, MA), 3 edition.

Busemeyer JA, Rapoport A (1988) Psychological models of deferred decision making. Journal of Mathematical Psychology 32:91–134.

Chan CW, Green LV, Lekwijit S, Lu L, Escobar G (2019) Assessing the impact of service level when customer needs are uncertain: An empirical investigation of hospital step-down units. Management Science 65(2):751–775.

de Vericourt F, Zhoug Y (2005) Managing response time in a call-routing problem with service failure. Operations Research 53(6):968–981.

Delasay M, Ingolfsson A, Bora K, Schultz K (2019) Load effect on service times. European Journal of Operational Research 279(3):673–686.

Edie L (1954) Traffic delays at toll booths. Journal of the Operational Research Society of America 2(2):107–138.

Fischbacher U (2007) z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10(2):171–178.

Forster AJ, Stiell I, Wells G, Lee AJ, Walraven CV (2003) The effect of hospital occupancy on emergency department length of stay and patient disposition. Academic Emergency Medicine 10(2):127–133.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: an empirical analysis of a maternity unit. Management Science 63(10):3147–3167.

Girotra K, Terwiesch C, K U (2007) Valuing r&d projects in a portfolio: Evidence from the pharmaceutical industry. Management Science 53(9):1452–1466.

Hathaway B, Kagan E, Dada M (2021) The gatekeeper's dilemma:"when should i transfer this customer?". Working Paper .

Hendriks A (2012) Sophie - software platform for human interaction experiments. University of Osnabrueck, Working Paper .

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. Manufacturing & Service Operations Management 15(2):263–279.

Ibanez M, Clark J, Huckman R, Staats B (2018) Discretionary task ordering – queue management in radiological services. Management Science 64(9):3971–4470.

Jaeker JAB, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. Management Science 63(4):1042–1062.

Janakiraman N, Meyer RJ, Hoch SJ (2011) The psychology of decisions to abandon waits for service. Journal of Marketing Research 48(6):970–984.

Kahneman D, Tversky A (1973) On the psychology of prediction. Psychological Review 80:237–251.

Kc CT Diwas S (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Management Science 55(9):1486–1498.

Kc DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. Manufacturing and Service Operations Management 14(1):50–65.

Keenan S, Sibbald W, Inman K, Massel D (1998) A systematic review of the cost-effectiveness of noncardiac transitional care units. Chest 113(1):172–177.

Kim S, Chan C, Olivares M, Escobar G (2015) Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. Management Science 61(1):19–38.

Kivetz R, Urminsky O, Zheng Y (2006) The goal-gradient hypothesis resurrected: Purchase acceleration, illusory goal progress, and customer retention. Journal of Marketing Research 43(2):39–58.

Kremer M, Minner S, Van Wassenhove L (2010) Do random errors explain newsvendor behavior? Manufacturing & Service Operations Management 12(4):673–681.

Kuntz L, Mennicken R, Scholtes S (2015) Stress on the ward: Evidence of safety tipping points in hospitals. Management Science 61(4):754–771.

Loch C, Terwiesch C (1999) Accelerating the process of engineering change orders: Capacity and congestion effects. Journal of Product Innovation Management 16(2):145–159.

Long EF, Mathews KS (2018) The boarding patient: Effects of icu and hospital occupancy surges on patient flow. Productions and Operations Management 27(2):2122–2143.

Lu Y, Musalem A, Olivares M, Schilkrut A (2013) Measuring the effect of queues on customer purchases. Management Science 59(8):1743–1763.

Mæstad O, Torsvik G, Aakvik A (2010) Overworked? on the relationship between workload and health worker performance. Journal of Health Economics 29(5):686–698.

Mas A, Moretti E (2009) Peers at work. American Economic Review 99(1):112–145.

Oliva R, Sterman JD (2001) Cutting corners and working overtime: Quality erosion in the service industry. Management Science 47(7):894–914.

Palley A, Kremer M (2014) Sequential search and learning from rank feedback: Theory and experimental evidence. Management Science 60(10):2525–2542.

Pitz GF, Reinhold H, Geller ES (1969) Strategies of information seeking in deferred decision making. Organizational Behavior and Human Performance 4:1–19.

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

Rabin M (2002) Inference by believers in the law of small numbers. The Quarterly Journal of Economics 775–816.

Rapoport A, Tversky A (1970) Choice behavior in an optimal stopping task. Organ. Behav. Human Performance 5:105–120.

Saad G, Russo JE (1996) Stopping criteria in sequential choice. Organizational behavior and human decision processes 67:258–270.

Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ (1998) Modeling and worker motivation in jit production systems. Management Science 44(12):1595–1607.

Schultz KL, McClain JO, Thomas JL (2003) Overcoming the dark side of worker flexibility. Journal of Operations Management 21:81–92.

Schweitzer M, Cachon G (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. Management Science 46(3):404–420.

Shafir E, Tversky A (1992) Thinking through uncertainty: Nonconsequential reasoning and choice. Cognitive Psychology 24:449–474.

Staats B, Gino F (2012) Specialization and variety in repetitive tasks: evidence from a japanese bank. Management Science 58(6):1141–1159.

Su X (2008) Bounded rationality in newsvendor models. Manufacturing and Service Operations Management 10(4):566–589.

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

Wang J, Zhou Y (2018) Impact of queue configuration on service time: Evidence from a supermarket. Management Science 64(7):2973–3468.

Recent ESMT Working Papers

	ESMT No.
The economics of dependence: A theory of relativity Hans W. Friederiszick, ESMT Berlin and E.CA Economics Steffen Reinhold, E.CA Economics and ECARES –Université Libre de Bruxelles	**21-02**
Beyond retail stores: Managing product proliferation along the supply chain Işık Biçer, SchulichSchool of Business, York University Florian Lücker, Cass Business School, City, University of London Tamer Boyaci, ESMT Berlin	**19-02 (R3)**
Effectiveness and efficiency of state aid for new broadband networks: Evidence from OECD member states Wolfgang Briglauer, Vienna University of Economics and Business (WU) Michał Grajek, ESMT Berlin	**21-01**
Contracting, pricing, and data collection under the AI flywheel effect Huseyin Gurkan, ESMT Berlin Francis de Véricourt, ESMT Berlin	**20-01 (R3)**
Informing the public about a pandemic Francis de Véricourt, ESMT Berlin Huseyin Gurkan, ESMT Berlin Shouqiang Wang, Naveen Jindal School of Management, The University of Texas at Dallas	**20-03 (R2)**