POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT

Vol. 23, No. 4, xxxx–xxxx 2014, pp. 704–722
ISSN 1059-1478 | EISSN 1937-5956 | 14 | 2304 | 0704

DOI 10.1111/poms.12089

Interruption and Forgetting in Knowledge-Intensive Service Environments

Craig M. Froehle, Denise L. White
Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, Ohio 45221-0130, USA,
James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio 45229-3026, USA, craig.froehle@uc.edu, denisel.white@cchmc.org

An increasing barrier to productivity in knowledge-intensive work environments is interruptions. Interruptions stop the current job and can induce forgetting in the worker. The induced forgetting can cause re-work; to complete the interrupted job, additional effort and time is required to return to the same level of job-specific knowledge the worker had attained prior to the interruption. This research employs primary observational and process data gathered from a hospital radiology department as inputs into a discrete-event simulation model to estimate the effect of interruptions, forgetting, and re-work. To help mitigate the effects of interruption-induced re-work, we introduce and test the operational policy of *sequestering*, where some service resources are protected from interruptions. We find that sequestering can improve the overall productivity and cost performance of the system under certain circumstances. We conclude that research examining knowledge-intensive operations should explicitly consider interruptions and the forgetting rate of the system's human workers or models will overestimate the system's productivity and underestimate its costs.

Key words: health care; services; interruptions; simulation
History: Received: September 2011; Accepted: April 2013 by Michael Pinedo, after 3 revisions.

1. Introduction

> The radiologist hung up the phone with an aggravated sigh. "What's the matter?" asked her colleague. "Well," she replied, "I've been trying for an hour to finish reading this exam, but every few minutes, I get paged or someone walks in here to ask a question. I then have to waste time retracing my whole train of thought just to get back to where I was before I got interrupted." "Ah, yes" nodded the other physician sympathetically, "it always seems to take twice as long to get anything done when you're on call and can't close your office door."

In 1991, Peter F. Drucker stated, "The single greatest challenge facing managers in the developed countries of the world is to raise the productivity of knowledge and service workers" (p. 69). However, despite the importance of knowledge-worker productivity (Davenport et al. 2002, Matson and Prusak 2010), ubiquitous connectivity, via technologies such as email and mobile telephony, and ever-rising expectations for service availability (Piccoli et al. 2009) have caused interruptions to become an increasingly common *barrier* to productivity[1] in professional and knowledge-intensive service environments (Bannister and Remenyi 2009, Seshadri and Shapira 2001). Knowledge-intensive service environments are those production systems where a significant portion of the work depends on concerted intellectual effort by trained professionals, consistent with, or at least a subset of, the "information-intensive services" domain (Apte et al. 2010) and the "white-collar" service domain (Hopp et al. 2009, Ramirez and Nembhard 2004).

Although interruptions have been studied analytically (e.g., Federgruen and Green 1986, Pang and Whitt 2009, Rao 1965, White and Christie 1958) and in various manufacturing and machine shop environments (Allwood and Lee 2004, Benkard 2000, Jaber and Bonney 2003), their operational effects on productivity have so far been largely overlooked in the context of knowledge-intensive services. The operations management literature broadly discusses interruptions (e.g., Seshadri and Shapira 2001, Speier et al. 2003), and some research has noted the effect of interruptions on service and/or knowledge-work environments, such as nursing (Tucker and Spear 2006), pharmacy (Flynn et al. 1999), surgery (Healey et al. 2006), and software engineering (Perlow 1999). But there is very little research that focuses primarily on interruptions as an integral production element in knowledge-work settings and also proposes and tests new policies for improved operational configurations (Seshadri and Shapira 2001 being a notable example).

In service industries, production often involves significant human labor. This is most true in knowledge-intensive environments, such as medicine, consulting, architecture, academia, etc., because comparatively little of the production process has been automated. This reliance on labor opens the door for a host of "human" issues in services, a useful overview of which is provided by Cook et al. (2002). From concerns about cognitive engagement to establishing the balance of control between the customer and employee to matching the service experience with customers' expectations, the customer's presence in and evaluation of the service creates many challenges (Bitner and Brown 2008, Ding et al. 2010). Whether these service encounters are "high-touch" or "high-tech" (technology mediated), understanding how both workers and customers function on a cognitive level during service co-production is vital to improving service operations (Bitner et al. 2000, Froehle and Roth 2004, Meuter et al. 2000).

Of central concern in this study is the possibility that this reliance on human labor for knowledge-intensive work creates a significant cost when it comes to interruptions. Unlike computers, humans tend to forget portions of the work that has already been completed when they stop working on a job partway through and return to it later. Finishing that job then requires additional re-work (or "re-learning") to re-assemble important details and return to the same level of information that had been achieved prior to the interruption so that work on the interrupted job can be resumed and completed (Bailey and McIntyre 2003, Finkenbinder 1913, Kher et al. 1999, Seshadri and Shapira 2001).

This study examines the operational effect of interruptions and the forgetting they induce in a knowledge-intensive service environment. While understanding this phenomenon may be important to a wide range of services, knowledge-intensive services are an appropriate first area of exploration given their susceptibility to interruptions. Using primary empirical data and simulation modeling, we test the policy we refer to as *sequestering*, or limiting the potential for interruption to a subset of available service resources. A sequestered server is protected from interruptions, which causes interruptions to flow to other servers not sequestered. The "closed door" work style of many professionals is *prima facie* evidence that this policy is widely used, yet there has been no concerted effort so far to understand its operational implications.

Therefore, in the presence of interruptions and forgetting-induced re-work, this research seeks insights into two primary questions. First, to what degree does forgetting affect the operational performance of a knowledge-intensive service production system? Second, does sequestering represent an effective approach to mitigating the effect of interruptions and, if so, how can we characterize an optimal sequestering policy?

The rest of this study is organized as follows. The next section presents relevant literature and forms testable hypotheses. The third section discusses the data collection and simulation methodologies used to test our hypotheses. Numerical results are presented in the fourth section. The final section offers conclusions based on those results, some research extensions and limitations of this study, and implications of the results for practice.

2. Literature Review and Theory Development

2.1. Interruptions

Interruptions are incidents or occurrences that impede regular work flow (Zellmer-Bruhn 2003) and are generally undesirable in most production systems. They can be caused by various factors, such as the nature of the work, physical layout, technology, etc. Interruptions come in many forms and have been broadly grouped under four categories: breaks, distractions, discrepancies, and intrusions (Jett and George 2003). An *intrusion*, the fourth category, is defined as "an unexpected encounter initiated by another person that interrupts the flow and continuity of an individual's work and brings that work to a temporary halt" (Jett and George 2003, p. 495). Unscheduled requests by co-workers or phone calls from customers, for example, could be considered intrusions in many cases. Intrusions impose the need to spend time with others on activities that may not be instrumental to completing the current (primary) job at hand. This study focuses exclusively on the impact of these *intrusion* interruptions due to their prevalence and their ability to be controlled (to some degree) through organizational policy, process design, and/or the layout of the work environment.

In manufacturing, interruptions are generally considered undesirable and they, or their effects, are typically minimized whenever possible by altering production schedules or inventory policies, instituting overtime, permitting back orders, adjusting batch sizes, preventative maintenance, and other operational levers (Groenevelt et al. 1992, Moinzadeh and Aggarwal 1997, Özbayrak et al. 2004, Raheja and Subramaniam 2002, Teyarachakul et al. 2011). The variability that interruptions induce is associated with waste and inefficiency, consistent with the "Theory of Swift, Even Flow" (Schmenner and Swink 1998), which summarily posits that uncertainty and variability in the operational environment degrades productivity.

However, pursuing a net reduction in interruptions is not always a viable tactic in many high-contact service environments, where an essential function of the service may be *accommodating* interruptions and responding to unscheduled requests. The presence of the customer in the production process necessitates flexible and responsive operations that can run efficiently even in the face of sporadic disruption (Sampson and Froehle 2006). One clear example is emergency health services, to which the concept of regular, smooth production is a largely foreign (and possibly irrelevant) notion. Because of this direct customer involvement, the results of interruptions, such as procedural justice violations (Brockner and Wiesenfeld 1996, Sparks and McColl-Kennedy 1998) and increased waiting time, can negatively affect customers' perceptions of service quality. Moreover, when service interruptions are generated by customers—a common occurrence—this creates the opportunity for what could be called the "interruption conundrum." On one hand the service provider risks alienating the customer currently being serviced to address the interruption because customers generally resent their service being pre-empted (at least without adequate justification or compensation). On the other hand, customers often have the *expectation* of being accommodated even though they may be causing an interruption to other customers. In service settings where both the current customer and the interrupting customer are present, the service provider faces potential negative consequences from *either* course of action; *both* tending to the interruption *and* ignoring it can damage customer relationships. This presents a significant dilemma for the service provider with no consistently obvious best policy. Clearly, it is important for us to better understand how interruptions can be managed operationally.

2.2. Forgetting

When a job is interrupted, completion of the job generally requires some amount of time longer than the job would have taken had it not been interrupted. This additional time arises from two sources. The first source is the interruption itself, which adds a length of time to the overall job duration equal to the duration of the interruption (this assumes that the interruption causes the worker to cease progress on the original job, which is consistent with the definition of "interruption" stated above). The second source of additional time is the extra work required by the worker to re-familiarize or remind himself of any details of the job that may have been forgotten (i.e., re-learning) during the interruption before further progress on the job can be made. This overall sequence has been referred to as the "learn, forget, and re-learn" process (Ash and Smith-Daniels 1999).

Forgetting is a complex cognitive phenomenon and has been examined from two primary perspectives: psychological recollection and industrial learning/forgetting curves. First, many psychology studies examining forgetting focus on the recollection of a singular past event, often in the form "I did such and such, at such and such time, in such and such place" (Sikström 2002). The effect of forgetting in these cases is typically measured by the quantity of details about the event lost from memory due to the passage of time.

The second perspective on forgetting, what we generally refer to here as the "learning/forgetting curve" perspective, is the one often taken by industrial engineering and operations researchers (e.g., Chiu et al. 2003, Smunt and Meredith 2000, Yelle 1979). This perspective sees interruptions as breaks in the repetition of a specific task over time (Teyarachakul et al. 2011). This interruption disrupts progress along the learning curve, impeding time/cost reductions, and complicating production planning (Bailey and McIntyre 2003, Carlson and Rowe 1976). Understanding interruptions conceptualized this way can help reduce task times (and thereby reduce overall costs) or help estimate future task times more accurately (thereby improving forecasting and/or project management efforts).

However, this study focuses on a slightly different cognitive scenario than either of these dominant perspectives. Our focus is the forgetting that occurs during an interruption *within a single job*. For example, a physician begins reviewing the history of a patient and the details of a recent procedure to make a decision regarding further treatment, but an interruption disrupts her progress on this job. The interruption delays completion of the job at hand and then, once the interruption has ended, forces the physician to re-trace her mental steps (adding additional work or re-work) to re-attain the level of knowledge about the initial job she had achieved before the interruption. Examples from other knowledge-work service settings could include an interruption while reviewing a lengthy legal contract or while analyzing a corporation's historical performance, product portfolio, and competitive landscape to forecast a future stock price target.

Although this scenario is somewhat similar to psychology's "recollection of past events" perspective described earlier, it differs in that it involves a complex set of details relevant to an *ongoing* job (very recent learning influences near-future learning and decision making) rather than the memory of something that had happened in its entirety some time ago.

Similarly, while the scenario in which we are interested involves many of the same elements that are of concern to the operational learning/forgetting curve researchers, we are examining the influence of interruption on a single (and possibly unique) job rather than on the repeated execution of a previously learned job. So, while this research draws upon these prior studies, it extends their lessons to a different forgetting context.

Many models of forgetting have emerged from these fields, all generally with the goal of parsimoniously predicting the amount of knowledge forgotten—a "forgetting rate"—based on the minimum number of essential parameters. Forgetting rates are generally described as a function of time; the more time that passes, the more is forgotten, with that relationship being modeled most commonly as a linear, exponential, logarithmic, or power function (Dar-El et al. 1995, Jaber and Bonney 1997, Jaber et al. 2003, Sikström 2002, Woltz and Shute 1995). Of these, the power function has produced the most consistently accurate representations of forgetting (Globerson et al. 1989, Jaber and Bonney 1997, Jaber et al. 2003, Sikström 2002). Teyarachakul et al. (2011) offer a helpful summary of forgetting models.

Taking the learning/forgetting curve perspective, which predominates in operations research, Bailey (1989) empirically examined the forgetting function and determined that forgetting is primarily determined by two factors: (a) the amount of learning prior to the interruption and (b) the duration of the interruption. Consistent with previous research, Bailey found that the power function again seemed to be most accurate in representing the forgetting phenomenon:

$$D = \alpha L \cdot \log(I). \tag{1}$$

where

$D =$ the additional time added to the job due to forgetting,

$\alpha =$ the forgetting rate,

$L =$ the amount of learning (i.e., progression along the learning curve before interruption), and

$I =$ the duration of the interruption.

The findings of Bailey (1989), as embodied in Equation (1), apply to our context to suggest that, as the duration of the interruption increases, so does the amount of job-specific information that is forgotten. In addition, the further along in the job the worker gets (i.e., the more job-specific information has been learned) prior to an interruption, the greater the opportunity for forgetting during an interruption. As knowledge worker tasks often involve "significant concentration and attention" (Davis and Naumann 1999, p. 344), this forgetting effect poses a potentially serious barrier to productivity.

When forgetting within a job occurs (due to an interruption or other causes), re-learning must occur for the job to be completed correctly. Researchers continue to debate the rate of re-learning, with some positing that the re-learning rate is the same as the original learning rate while others argue that it may be different (Bailey and McIntyre 2003, Jaber et al. 2003). As there is no resolution yet on this issue, we assume here that the re-learning rate is equivalent to the original learning rate (i.e., no additional penalty is paid nor efficiency gained by virtue of having to re-learn previously learned information) and leave the relaxation of that assumption to future work.

2.3. De-coupling and Sequestering

In many service environments, workers are tasked with a mixture of front-office activities (involving direct contact with customers) and back-office activities (no direct contact with customers). Such environments are typically referred to as "mixed services" (Chase 1981). One common tactic to handling customer-induced interruptions in these situations is to *de-couple* the front-office tasks from the back-office tasks by separating them procedurally, if not also physically/geographically (Metters and Vargas 2000, Metters and Verma 2008). This tends to isolate the back-office workers and reduces their exposure to customer-induced interruptions and variability (Sampson and Froehle 2006).

This concept draws upon the strategic notion of the "focused factory" (Skinner 1974), which has been suggested as an approach to increase quality and reduce the costs associated with developing multiple competencies in an organization. Although the concept originated in manufacturing, focus has also been validated as an operations strategy in *service* organizations, including health care (Hyer et al. 2009). Thus, it seems reasonable that a *focused* service production system could be designed to minimize customer interruptions for part of the organization while concentrating those interruptions that must be accommodated within a different part of the organization. Modern call centers, which concentrate customer service requests into that designated subset of the organization, are an example of focus in services and an alternative to the "mixed" services model.

In the radiology function of a hospital, as an example of a knowledge-work environment, a typical arrangement is for radiologists to spend much of their time reading exams (i.e., an efficiency-oriented, back-office activity) while simultaneously handling interruptions created by other medical staff, patients, outside physicians, etc. (i.e., a customer-intensive, front-office activity); a "mixed services" model.

Applying the concepts of focus and de-coupling to this environment would suggest that *sequestering* one or more of the radiologists (resources) away from direct contact with customers may help reduce the productivity degradation attributed to interruptions. Enacting such a policy creates a buffer against customer-induced interruptions for those temporarily sequestered radiology resources, thereby hypothetically increasing the efficiency with which they can process exams.

If we assume that interruption demand is exogenous, sequestering would shift that interruption demand to un-sequestered resources. This potentially has two simultaneous effects: (a) reducing the productivity (in terms of routine work completed per-unit time) for those non-sequestered resources and (b) increasing the queue length/waiting time for interruptions (because there would be fewer resources available to process them).

Considering the range of policies that could involve sequestering, at one extreme, there would be no sequestering; all resources would have equal probability of being interrupted at any time. At the other extreme, sequestered resources would be entirely protected from interruption (i.e., they would be completely isolated from customers' unscheduled requests). As it may be difficult for a service worker to be entirely sequestered away from customer contact, a sequestering policy could vary depending on the severity/urgency of the need to interrupt the sequestered resource. For example, one such policy might limit interruptions of a sequestered worker to just certain types of high importance (a subset of all possible interruptions). This type of *variable* sequestering has not, to our knowledge, been examined before in the literature.

2.4. Hypotheses

In investigating these ideas, if we assume that *interruptions are disruptions to routine production work*, and that a single set of servers is having to process both kinds of jobs (routine and interruption), the observations outlined above can be summarized in the following hypotheses, all stated in the alternative form:

> Hypothesis 1: Sequestering decreases average flow time for routine production jobs.

> Hypothesis 2: Sequestering decreases average waiting time for routine production jobs.

> Hypothesis 3: Sequestering increases average flow time for interruption jobs.

> Hypothesis 4: Sequestering increases average waiting time for interruption jobs.

We define "flow time" as the duration a job takes from the moment it enters the system to when it is completed and exits the system, and we define "waiting time" as the total time a job spends in the system not being actively processed; waiting time is a subset of flow time, but only for those jobs that experience waiting. While these hypotheses may bear out in testing, it would still not be clear if the throughput improvement for production work would outweigh the reduced service level provided to interruptions. To assess that trade-off, one proxy measure might be the total amount of time required to process a given workload (i.e., mix of production work and interruptions). If the overall time required to handle the entire workload decreases, then the system could be argued as being more efficient. This leads us to an additional hypothesis, again stated in the alternative form:

> Hypothesis 5: The total shift duration—the time required to process a given quantity of jobs (the sum of routine production work and interruptions)—will decrease as sequestering increases.

However, time measures alone may not represent the best or most important metric by which to judge the merits of a particular sequestering policy. Interruptions may be allowed to pre-empt routine work for a variety of reasons, but the most obvious, and perhaps the most compelling, is that the per-unit time "cost" of making the interruption wait is higher than the per-unit time "cost" of making the routine production work wait. In other words, delaying the interruption is costlier in some way than delaying the routine job currently at hand. Of course, these "costs" may not always be purely monetary.

Returning to our previous example, a radiologist may be reading an exam (i.e., performing routine production work) when she is interrupted by an emergency department (ED) physician with a question about a current trauma case. Ignoring that ED physician and making him wait (forcing him to be idle) is expensive by itself (i.e., the loss of the ED doctor's time at his hourly rate), but possibly no more expensive than making a *different* physician wait longer for the results on the routine exam being read at that moment. However, delaying the ED physician *also* potentially holds up an entire ED team (e.g., fellows, nurses, technologists, etc.) *and* possibly exposes the hospital to additional medicolegal risk as a result of delayed care to a critically ill patient. The need to minimize the sum of all those costs associated with forcing expensive resources to be idle (in part) motivates the radiologist to pre-empt the exam in process (routine production work) and attend to the ED physician's request (interruption).

Therefore, taking the cost of time into consideration and assuming that interruptions are at least as costly as routine production work (never less, or they would be unlikely to warrant pre-emption of work in progress) and Hypotheses 3 and 4 are true, we would expect the value of a sequestering policy, which reduces the service level provided to interruptions, to be dependent on how much more costly it is to delay interruptions as compared to exams. With this in mind, our final hypothesis is as follows:

> Hypothesis 6: As the ratio of the cost of delaying interruptions to the cost of delaying routine production work increases, overall system costs will increase with sequestering.

Using an empirically based discrete-event simulation model, we test each of the first five hypotheses under various levels of forgetting rate, sequestering, and interruption arrival rate. Those numerical results are then further analyzed by adding cost information to test Hypothesis 6 and derive an optimal sequestering policy.

3. Methods

This research employed different methodologies in two stages. In the first stage, we collected empirical data for inter-arrival times and durations for routine jobs and interruptions. In the second stage, we constructed a simulation model using the empirical data from the first stage as inputs. Both methodologies are detailed below.

3.1. Study Environment and Data Collection

To estimate some of the basic operating parameters related to production work and interruptions within a knowledge-intensive service environment, we observed and gathered data from the radiology department of a large, Midwestern teaching hospital. While a radiology department is certainly not generalizable to all knowledge-intensive service environments, it does serve well as an appropriate context to use as a first step in examining the operational effects of interruptions and in assessing the potential usefulness of sequestering.

The work performed by this radiology department's central reading room consists primarily of two activities: (a) analyzing various types of radiology exams (x-rays, CT scans, etc.), which is considered routine production work, so that downstream physicians can complete their decision making, and (b) accommodating unscheduled requests (e.g., answering questions from other medical staff, taking phone calls from referring physicians, etc.), which are considered interruptions. The busiest time in the central reading room is during the evening (approximately 5 PM–9 PM) each day, where work is typically performed by a team of two physicians. This team processes exams as they are generated by the hospital's ED and various satellite (outpatient) locations, often building up a temporary backlog during especially busy times. This team is also still subject to a variety of intrusion interruptions as many health-care organizations are 24/7 operations and patients do not cease being ill at 5 PM.

The typical non-emergent radiology exam reading process we have modeled can be summarized as such: when an exam arrives at the reading room to be read, it is either picked up by an idle radiologist or added to the queue of waiting exams. Exams can be read by either of the two radiologists and are processed first come, first served (FCFS). At 10 PM each evening, the outpatient centers stop generating new exams and the two radiologists continue to read queued exams until all exams are processed, at which point they are sent home or allowed to attend to other duties. Any routine requests that arrive after 10 PM are routed to a separate overnight radiologist outside of the work system being modeled.

While the radiologists are processing exams in the central reading room, they are frequently interrupted by other physicians, nurses, radiology technologists, phone calls, and various other events. Interruptions can and often do pre-empt an exam currently being read; this causes the radiologist to set aside the exam currently being read to concentrate on the interruption request. Once the interruption has been tended to, the radiologist is free to continue processing the pre-empted exam. This is not unlike the situation faced by many customer-accessible knowledge workers who split their time between addressing customers' immediate needs and working on other tasks.

The data to be used as inputs to the simulation were collected from this after-hours radiology reading room environment. In-person monitoring of the reading room for 40 hours over a period of 2 weeks (the four busiest hours each evening, five evenings each week) enabled us to collect detailed interruption data that could not be gleaned from routine operational or patient records. These observational data provided inter-arrivals and durations for various types of interruptions. Interruptions used for the analysis were limited to those meeting the definition of *intrusions* previously described (interruptions that cannot be controlled by the physician). This precluded breaks in workflow initiated by the radiologists themselves (e.g., bathroom breaks, meals, outgoing phone calls, etc.).

To estimate the inter-arrival and reading (processing) times for exams, we also collected relevant work papers during the same time frame. This archival data source provided time stamps for important points throughout the radiology exam reading process (Halsted and Froehle 2008). Distributions were fit to these empirical data (see Table 1 for distributions and representative statistics) and employed as parameter inputs to the simulation model, which is described next.

3.2. Simulation Model

We constructed a discrete-event simulation model representative of the workflow and decision making present in the central reading room during evening hours, as described above. In our model (see Appendix), after-hours (evening) radiology exams begin arriving at the central reading room at 5 PM and stop arriving at 10 PM (when the outpatient locations close for the day). Radiology exams generated by the ED are handled by a dedicated ED radiologist starting at 10 PM, so those exams are no longer routed to the central reading room after 10 PM. However, at 10 PM, there may still be a backlog of exams waiting to be processed by the reading room radiologists, so the shift typically extends somewhat past 10 PM. The total duration of this work time—from 5 PM. until the time when all interruptions and exams have been processed—is what we refer to here as "shift duration."

Three variables drive the scenarios that were simulated: forgetting rate α (five levels), sequestering level θ (five levels), and interruption arrival rate λ_i (two levels). At forgetting rates greater than zero, the added re-work due to re-learning is calculated based on Equation (1). When an exam reading is interrupted, we determine the new remaining reading time R by

$$R = T + D, \tag{2}$$

where

$T =$ the time remaining out of the original total job time (or out of the *updated* total job time if the job is interrupted on multiple occasions), and

$D =$ the added time due to re-learning/re-work (see Equation (1)

Example: An exam initially requires an uninterrupted processing time of 10 minutes. Six minutes into the job, the physician is interrupted for 5 minutes. If $\alpha = 1$, the *updated* remaining time required to process this exam (once the physician continues working on it) would be $4 + 1 \bullet 6 \bullet \log(5) = 8.2$ minutes. Assuming no further interruptions to this exam, the final total work time required to process the exam would be 14.2 minutes. If we also include the 5-minute interruption, the "completion time" (Gaver 1962) of the exam would be 19.2 minutes. In this example, if re-work is not considered (when $\alpha = 0$), the remaining processing time after the interruption would be just 4 minutes.

Note that, consistent with Teyarachakul et al. (2011), we limit the value of D so that in no case would the amount of re-work created by a single interruption exceed the amount of work (learning) completed on a job prior to the interruption (e.g., if an interruption occurs 2 minutes into a job, the re-work associated with that interruption will not exceed 2 minutes regardless of the length of the interruption). In other words, an interruption cannot remove an amount of learning about a job *greater* than what a worker had achieved from the start of the job to the point of the interruption; forgetting everything one has learned about the job is the most forgetting an

Table 1 Summary Statistics and Fitted Distributions for Simulation Input Data

	Exam durations	Exam inter-arrivals	Interruption durations	Interruption inter-arrivals
N	204	210	201	200
Mean	5.18	5.97	1.79	9.03
SD	8.6	6.95	2.11	11
Min	0	0	0	0
Max	41	46	14	62.8
Distribution	Beta	Gamma	Weibull	Gamma
Expression	41•BETA(0.236, 1.36)	GAMM(11.9, 0.502)	WEIB(1.79, 1)	GAMM(19.1, 0.473)
Square error	.006	.004	.003	.003
Chi-square test				
Number of intervals	6	5	6	6
Degrees of freedom	3	2	3	3
Test statistic	12.6	8.76	6.04	13.7
p-value	0.006	0.014	0.115	< 0.005
Kolmogorov–Smirnov test				
Test statistic	0.234	0.060	0.060	0.411
p-value	<0.01	>0.15	>0.15	<0.01

interruption can cause. We considered five levels of forgetting rate: $\alpha = 0$, .25, .5, .75, and 1, with values above 1 representing extreme levels of individual forgetting that would be unlikely to occur in practice.

The second variable influencing the simulation results is sequestering (θ), or the degree to which the sequestered resource is protected against interruptions. As modeled here, sequestering determines the proportion of interruptions directed *away from* the sequestered resource and toward the un-sequestered resource. For example, "sequestering $\theta = .8$" indicates that 80% of all interruptions that would have been directed to the sequestered server (physician) will instead be handled by the other server (i.e., resulting in a 90%/10% allocation of interruptions overall). Five levels of sequestering were applied to one of the two-server resources: $\theta = 0$ (no protection from interruptions), .5, .8, .9, and 1.0 (complete protection). The values were concentrated in the upper half of the range due to the possibility that sequestering might have a non-linear effect due to the queueing characteristics of the model.

Finally, the third variable driving the scenarios is the arrival rate of interruptions (λ_i). We desired to determine if the effects of the forgetting rate (α) and sequestering (θ) variables were dependent on (interacted with) the interruption rate in such a way that our resulting policy recommendations were only generalizable to environments with similar interruption levels. As the environment from which we obtained our empirical service data was not exceptionally busy at the original arrival rate of interruptions (λ_i)—overall traffic intensity was ~0.53—we also simulated all scenarios at twice that rate ($2\lambda_i$).

The simulation model involves two radiologist resources (one attending ["att"] and one fellow ["fel"], named such solely for modeling convenience as only the "att" resource is sequestered). Both resources are assumed to be of equal skill (i.e., equal service rates), have equal forgetting rates, and each resource can process both examinations (routine production work) and interruptions. We also assume that examinations and interruptions arrive independently, according to their respective inter-arrival distributions. Examination and interruption processing times are independent and are based on their respective duration distributions.

In the modeled workflow, all incoming exams wait in a single, common first-in-first-out (FIFO) queue for the next available resource. Incoming interruptions are randomly and immediately directed to one of the two resources, with that assignment being either unbiased (in the case of zero sequestering) or biased toward the fellow (the un-sequestered server) as determined by the sequestering level. If the target resource is idle, the interruption is processed immediately. If the chosen resource is busy with an exam, the exam is immediately pre-empted in order for the interruption to be processed. Interruptions cannot pre-empt other interruptions; in such a case, the incoming interruption waits in a FIFO queue dedicated to its assigned resource. A server returns to processing its pre-empted exam immediately after the interruption has been served (unless there is another interruption waiting in that server's interruption queue). An interrupted exam can be completed only by the resource that started processing it—there are no hand-offs of exams.

In the work system modeled here, exam requests and interruptions enter the system for 300 minutes (representing the period from 5 PM to 10 PM during a single workday) and then stop arriving (for reasons previously described). The simulated reading room continues to process all queued exams and interruptions until none remain, at which point the simulation run ends (the physicians go home for the evening). With the full-factorial experimental design used here, each of the 50 scenarios (five levels of forgetting rate, five levels of sequestering, and two levels of interruption arrival rate) was subjected to 1000 replications to ensure appropriately small 95% confidence intervals (Kelton et al. 2004). All input data were synchronized across scenarios to ensure that performance differences were not due to sampling. The complete simulation logic diagram is provided in the Appendix.

4. Results

4.1. Hypothesis Testing

Quantitative results from the simulation and our hypothesis testing are given below. Table 2 presents numerical results for time-based outcome measures across a selected subset of simulated conditions, and Table 3 provides the results of paired *t*-tests comparing corner point (α, θ = 0,0; 0,1; 1,0; and 1,1) outcome measures. Note that Table 3 does not include comparisons "across the diagonal," as comparing scenarios where both variables (α and θ) change simultaneously did not contribute additional insight.

Our first hypothesis stated that average exam flow times should decrease as sequestering (θ) is increased. Figure 1 shows exam flow times plotted against sequestering level (θ) and forgetting rate (α) for both levels of interruption arrival rate (λ_i). Table 2 provides representative numerical results and Table 3 offers up statistical tests of corner point comparisons. At the zero re-work level (i.e., when $\alpha = 0$, or when forgetting-induced re-work was not considered in the simulation), exam flow time remained essentially unchanged at roughly 16 minutes across all sequestering scenarios at λ_i, and 21 minutes at $2\lambda_i$. However,

Table 2 Numerical Simulation Results (Abbreviated)

Interruption arrival rate	Forgetting rate (α)	Sequestering level (θ)	Mean exam flow time	Mean exam wait time	Mean interruption flow time	Mean interruption wait time	Mean total shift duration	Total exam flow time	Total interruption flow time
λ_i	0	0	15.99 (.56)	9.17 (.51)	2.12 (.03)	0.35 (.02)	320.74 (1.16)	871.56 (39.00)	75.87 (1.55)
		0.8	15.96 (.56)	9.14 (.51)	2.43 (.04)	0.65 (.03)	320.95 (1.16)	870.06 (38.84)	87.34 (2.08)
		1	15.95 (.56)	9.13 (.51)	2.60 (.04)	0.82 (.03)	321.09 (1.16)	869.61 (38.87)	93.64 (2.32)
	0.5	0	17.82 (.68)	10.72 (.62)	2.12 (.03)	0.35 (.02)	323.24 (1.28)	976.57 (46.94)	75.91 (1.55)
		0.8	17.35 (.64)	10.30 (.59)	2.43 (.04)	0.65 (.03)	322.89 (1.25)	949.46 (44.78)	87.37 (2.08)
		1	17.09 (.63)	10.06 (.57)	2.60 (.04)	0.82 (.03)	322.71 (1.24)	934.29 (43.61)	93.67 (2.32)
	1	0	19.71 (.81)	12.34 (.76)	2.12 (.03)	0.35 (.02)	325.83 (1.42)	1085.07 (56.18)	75.94 (1.55)
		0.8	18.79 (.75)	11.53 (.7)	2.43 (.04)	0.65 (.03)	324.97 (1.36)	1032.48 (52.08)	87.40 (2.08)
		1	18.20 (.70)	11.00 (.65)	2.60 (.04)	0.82 (.03)	324.47 (1.34)	998.54 (48.88)	93.70 (2.32)
$2\lambda_i$	0	0	21.02 (.81)	14.20 (.76)	2.48 (.03)	0.70 (.02)	325.57 (1.41)	1155.11 (55.85)	172.22 (3.15)
		0.8	21.01 (.80)	14.19 (.75)	3.26 (.06)	1.48 (.05)	325.67 (1.41)	1153.71 (55.44)	228.63 (5.43)
		1	20.94 (.80)	14.12 (.75)	3.73 (.07)	1.95 (.07)	325.66 (1.41)	1149.92 (55.16)	262.84 (6.96)
	0.5	0	25.03 (1.03)	17.75 (.98)	2.48 (.03)	0.70 (.02)	330.21 (1.63)	1381.02 (70.47)	172.25 (3.15)
		0.8	23.73 (.95)	16.57 (.90)	3.26 (.06)	1.48 (.05)	328.95 (1.56)	1305.50 (64.76)	228.69 (5.42)
		1	22.84 (.89)	15.77 (.84)	3.73 (.07)	1.95 (.07)	328.30 (1.52)	1256.13 (61.32)	262.88 (6.96)
	1	0	29.36 (1.28)	21.63 (1.22)	2.48 (.03)	0.70 (.02)	335.54 (1.90)	1625.56 (86.28)	172.28 (3.15)
		0.8	26.51 (1.10)	18.95 (1.05)	3.26 (.06)	1.48 (.05)	332.53 (1.75)	1456.67 (74.61)	228.69 (5.43)
		1	24.33 (.97)	17.05 (.92)	3.73 (.07)	1.95 (.07)	330.91 (1.63)	1340.56 (66.56)	262.91 (6.96)

All values are means of 1000 iteration means, expressed in minutes; estimate half-widths shown in parentheses; mean flow time is overall jobs (exams or interruptions), whereas mean wait time is based only on that subset of jobs subjected to waiting. Total exam flow time and total interruption flow time are sums of flow times for all jobs of that type. Table 2 does not show results for certain values of α (0.25 and 0.75) and θ (0.5 and 0.9) in the interest of space; complete numerical results are available upon request from the authors.

Table 3 Single-Variable, Paired *t*-Test Comparisons of Outcome Measures

Comparison of scenario (α, θ)		$\alpha = 0$, $\theta = 0$	0, 1	0, 0	1, 0
to scenario (α, θ)		$\alpha = 1$, $\theta = 0$	1, 1	0, 1	1, 1
Interpretation		Forgetting increases to 1, no sequestering	Forgetting increases to 1, max sequestering	Sequestering increases to 1, no forgetting	Sequestering increases to 1, max forgetting
λ_i	Mean exam flow time	**8.43 (<.001)**	**5.01 (<.001)**	0.1237 (0.902)	**3.82 (<.001)**
	Mean exam wait time	**8.39 (<.001)**	**4.71 (<.001)**	1.3 (0.195)	**2.95 (0.003)**
	Mean interruption flow time	0.46 (0.644)	0.35 (0.729)	**9.41 (<.001)**	**9.41 (<.001)**
	Mean interruption wait time	0.6378 (0.524)	0.1844 (0.846)	**10 (<.001)**	**9.71 (<.001)**
	Mean total shift duration	**4.3435 (<.001)**	**2.3682 (0.018)**	0.0007 (0.999)	**1.9941 (0.046)**
	Total exam flow time	**6.9748 (<.001)**	**4.0378 (<.001)**	0.089 (0.929)	**3.2019 (0.001)**
	Total interruption flow time	0.3376 (0.735)	0.4038 (0.686)	**7.0115 (<.001)**	**7.0734 (<.001)**
$2\lambda_i$	Mean exam flow time	**14.2628 (<.001)**	**5.67 (<.001)**	0.413 (0.678)	**9.577 (<.001)**
	Mean exam wait time	**13.6486 (<.001)**	**5.401 (<.001)**	1.262 (0.207)	**8.078 (<.001)**
	Mean interruption flow time	1.848 (0.065)	0.924 (0.356)	**19.868 (<.001)**	**19.868 (<.001)**
	Mean interruption wait time	1.663 (0.096)	0.462 (0.644)	**19.989 (<.001)**	**19.418 (<.001)**
	Mean total shift duration	**9.901 (<.001)**	**4.19 (<.001)**	0.053 (0.958)	**5.953 (<.001)**
	Total exam flow time	**11.673 (<.001)**	**4.629 (<.001)**	0.325 (0.745)	**7.819 (<.001)**
	Total interruption flow time	1.099 (0.272)	0.638 (0.524)	**14.381 (<.001)**	**14.193 (<.001)**

Values shown are t (p-value); comparisons significantly different at $p < .05$ are shown in bold.

as higher levels of forgetting (and, thus, re-work) were considered, exam flow time decreased significantly as sequestering increased (per parameter estimate half-widths, $p = .05$, and paired t-tests). Thus, Hypothesis 1 was supported, but only when forgetting-induced re-work was specifically considered (i.e., $\alpha > 0$), and the effect was more pronounced in the busier system ($2\lambda_i$).

Hypothesis 2 suggested that average waiting time for exams would decrease with increased sequestering. Figure 2 and Table 2 show exam waiting times against sequestering level and forgetting rate, again for both levels of interruption arrival rate. These results indicate that, when forgetting-induced re-work was zero, exam waiting times remain essentially unchanged at just over 9 minutes for λ_i and just over 14 minutes for $2\lambda_i$. However, like the results for Hypothesis 1, when forgetting/re-work increased above zero, the average exam waiting time decreased in both levels of interruption arrival rate and the mag-

Figure 1 Mean Exam Flow Times

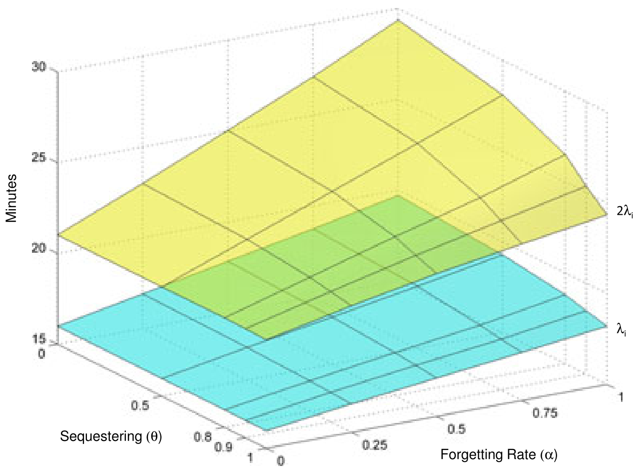


Figure 2 Mean Exam Wait Times

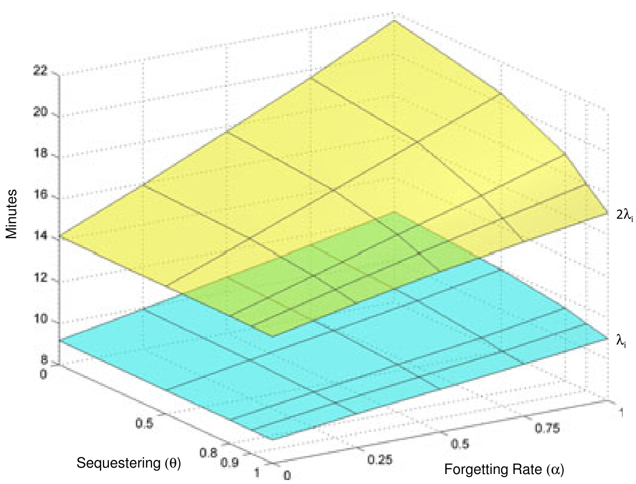
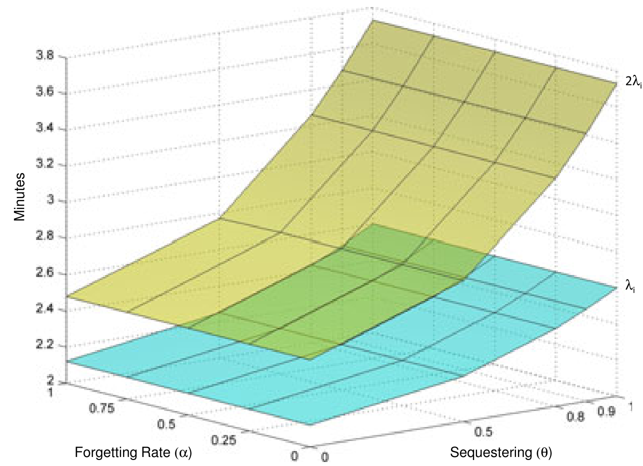


Figure 3 Mean Interruption Job Flow Times



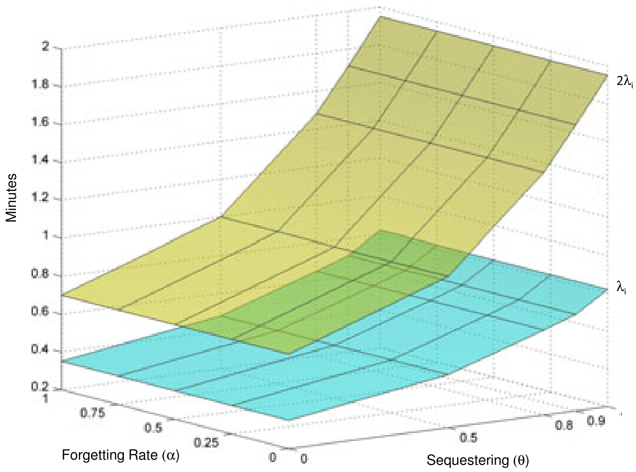
nitude of the reduction increased as the arrival rate of interruptions doubled (see Table 3). These results support Hypothesis 2, but only if re-work is explicitly included in the simulation (i.e., $\alpha > 0$); otherwise, if $\alpha = 0$, sequestering has no effect on exam waiting time.

Our third hypothesis posited that average flow time for interruptions would increase with increased sequestering. Figure 3 and Table 2 show interruption flow times against sequestering level and forgetting rate for both levels of interruption arrival rate. The results suggest that flow times increased significantly as sequestering increased, regardless of forgetting rate. And, as would be expected, when interruption arrival rate doubles, interruption flow time also increases (see Table 3). These results offer unconditional support for Hypothesis 3. Note that the forgetting rate does not affect interruption flow time, as, in our model, interruptions are never pre-empted; therefore, forgetting does not affect the processing of interruptions.

Our fourth hypothesis predicted that average interruption waiting times would increase with higher levels of sequestering. The average waiting times (overall interruptions that experienced a wait, regardless of which server processes them) are plotted against forgetting rate and sequestering level and are shown in Figure 4. From that chart, as well as Table 2 and Table 3, we notice that the average waiting time for all interruptions increased slightly, but significantly, as sequestering increased, regardless of forgetting rate. This trend was repeated and amplified in the higher interruption arrival rate condition. Hypothesis 4, therefore, also appears to be unconditionally supported.

Hypothesis 5 expected that total shift duration, or the time from the start of the shift to the time when all work is completed, would decrease as sequestering increased. If supported, this would suggest that the net time effect of sequestering on the conditions tested is beneficial in that it reduces the overall amount of work and decreases the time required to complete any given quantity of incoming (exogenous) work. Figure 5 shows the total shift duration plotted against forgetting rate and sequestering level for both levels of interruption arrival rate, with Table 2 again providing numerical results and statistical tests in Table 3. For the baseline interruption arrival rate (λ_i), the net reduction in shift duration due to sequestering just meets the $p < .05$ threshold. The results for the higher traffic intensity case ($2\lambda_i$) are more pronounced: Sequestering can slightly, but significantly, reduce the total time required to complete a shift's workload. As before, operational benefits of sequestering are not observed if forgetting-induced re-work is excluded from the model (i.e., at $\alpha = 0$, sequestering has no effect on re-work). Therefore, Hypothesis 5 is conditionally supported.

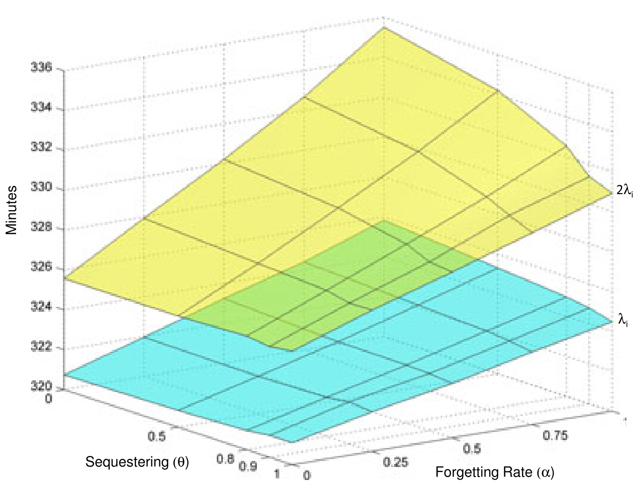
Figure 4 Mean Interruption Job Wait Times



Our final Hypothesis 6 posits that, as the ratio of interruption time cost to production work time cost increases, overall system costs will increase with sequestering. To test Hypothesis 6, the above simulation results were augmented with a range of values for cost variable C, the ratio of the per-unit cost of time associated with delaying interruptions (C_i) to the per-unit cost of time associated with delaying routine production work, or, in this case, exams (C_x). Therefore, $C = C_i/C_x$.[2] The average total cost was calculated for each of the five levels of forgetting (α), five levels of sequestering (θ), and two levels of interruption arrival rate (λ_i). Those results are presented in Table 4.

From Table 4, the most noteworthy observation is that, at any given cost ratio C and forgetting level α, cost minimization only occurs at one of the two extreme points of θ (0% and 100%); sequestering at a level *between* 0% and 100% is never optimal.[3] Therefore, only results relating to the two extreme values of θ (0 and 1) will be considered henceforth.

Figure 5 Total Shift Duration



A second important observation from Table 4 is that, while a "sequester vs. no sequester" (i.e., θ = 0% vs. θ = 100%) comparison for any given combination of C, α, and λ_i is straightforward, policy comparisons *across* values of C or α are not due to differentials in baseline (at θ = 0%) costs. Therefore, we standardize the total cost for each sequestering (θ = 100%) condition as its ratio to the corresponding "no sequestering" (θ = 0%) total cost. For example, the standardized total cost for a sequestering policy at (C = 10, α = .75, λ_i) is 1902.78/1788.12, or 1.06. This indicates that, under these conditions, the total costs of sequestering would be 106% of the total cost associated with no sequestering, so *not* sequestering would be preferred (i.e., less costly) in that situation. The resulting standardized sequestering costs are plotted against cost ratio C in Figure 6a–e.

In Figure 6a–e, C'_{λ} and $C'_{2\lambda}$ indicate the cost ratio transition points at which it is no longer beneficial to sequester; at cost ratios below C', sequestering is preferable (i.e., results in lowest total costs), and at cost ratios above C', *not* sequestering is preferable. Figure 7 plots these transition points, thereby showing the resulting preferred sequestering policies for the environments/scenarios modeled.

Hypothesis 6 posited that as C increases, total costs will rise with sequestering. All the cost data and associated figures clearly illustrate that this hypothesis is supported. In Figure 7, this relationship can be observed in action by the fact that at any level of α above zero, as C increases, the optimal policy shifts from 100% sequestering to 0% sequestering.

4.2. Sensitivity Analyses

To understand our results better and to improve the generalizability of the system as modeled, we extended our results through a series of sensitivity analyses, or limited simulations/calculations that test additional parameters or parameter values outside the full-factorial design described above.

First, we tested an extension of our workflow rules that changes how interruptions are handled. The baseline model (above) randomly assigns each incoming interruption to one of the two resources regardless of their state (idle or busy). This is simple and not unlike the observed environment from where our empirical data were obtained. However, there may be situations where an incoming interruption is routed opportunistically first to an idle resource, if any (or randomly chosen if both are idle or both are busy), so we tested that policy as well. Those results revealed no significant change to exam flow/wait times, but tiny improvements to interruption flow/wait times and overall clinic duration. These improvements were not large enough to be managerially meaningful, and they did not force us to qualify our earlier results,

Table 4 Numerical Results—Average Total Clinic Cost

		Interruption Arrival Rate									
		λ_i (Sequestering level [θ])					$2\lambda_i$ (Sequestering level [θ])				
Cost ratio (C)	Forgetting level (α)	0	0.5	0.8	0.9	1	0	0.5	0.8	0.9	1
$C = 1$	0.00	947.43	950.83	957.40	960.03	963.25	1327.33	1347.29	1382.34	1397.16	1412.76
	0.25	999.73	999.00	997.98	997.78	997.79	1437.82	1445.65	1458.56	1463.54	1469.25
	0.50	1052.48	1044.81	1036.83	1032.49	1027.96	1553.27	1545.17	1534.19	1525.31	1519.01
	0.75	1104.75	1091.60	1074.15	1065.50	1059.57	1669.97	1644.55	1606.22	1584.02	1561.56
	1.00	1161.01	1140.93	1119.88	1105.01	1092.24	1797.84	1746.77	1685.36	1642.05	1603.47
$C = 2$	0.00	1023.30	1031.62	1044.74	1050.57	1056.89	1499.55	1541.05	1610.97	1642.21	1675.60
	0.25	1075.61	1079.79	1085.34	1088.35	1091.45	1610.06	1639.44	1687.22	1708.61	1732.13
	0.50	1128.39	1125.61	1124.20	1123.06	1121.63	1725.52	1738.97	1762.88	1770.40	1781.89
	0.75	1180.68	1172.43	1161.55	1156.09	1153.26	1842.24	1838.37	1834.91	1829.12	1824.45
	1.00	1236.95	1221.76	1207.28	1195.61	1185.94	1970.12	1940.60	1914.05	1887.16	1866.38
$C = 3$	0.00	1099.17	1112.41	1132.08	1141.11	1150.53	1671.77	1734.81	1839.60	1887.26	1938.44
	0.25	1151.49	1160.58	1172.70	1178.92	1185.11	1782.30	1833.23	1915.88	1953.68	1995.01
	0.50	1204.30	1206.41	1211.57	1213.63	1215.30	1897.77	1932.77	1991.57	2015.49	2044.77
	0.75	1256.61	1253.26	1248.95	1246.68	1246.95	2014.51	2032.19	2063.60	2074.22	2087.34
	1.00	1312.89	1302.59	1294.68	1286.21	1279.64	2142.40	2134.43	2142.74	2132.27	2129.29
$C = 4$	0.00	1175.04	1193.20	1219.42	1231.65	1244.17	1843.99	1928.57	2068.23	2132.31	2201.28
	0.25	1227.37	1241.37	1260.06	1269.49	1278.77	1954.54	2027.02	2144.54	2198.75	2257.89
	0.50	1280.21	1287.21	1298.94	1304.20	1308.97	2070.02	2126.57	2220.26	2260.58	2307.65
	0.75	1332.54	1334.09	1336.35	1337.27	1340.64	2186.78	2226.01	2292.29	2319.32	2350.23
	1.00	1388.83	1383.42	1382.08	1376.81	1373.34	2314.68	2328.26	2371.43	2377.38	2392.20
$C = 5$	0.00	1250.91	1273.99	1306.76	1322.19	1337.81	2016.21	2122.33	2296.86	2377.36	2464.12
	0.25	1303.25	1322.16	1347.42	1360.06	1372.43	2126.78	2220.81	2373.20	2443.82	2520.77
	0.50	1356.12	1368.01	1386.31	1394.77	1402.64	2242.27	2320.37	2448.95	2505.67	2570.53
	0.75	1408.47	1414.92	1423.75	1427.86	1434.33	2359.05	2419.83	2520.98	2564.42	2613.12
	1.00	1464.77	1464.25	1469.48	1467.41	1467.04	2486.96	2522.09	2600.12	2622.49	2655.11
$C = 6$	0.00	1326.78	1354.78	1394.10	1412.73	1431.45	2188.43	2316.09	2525.49	2622.41	2726.96
	0.25	1379.13	1402.95	1434.78	1450.63	1466.09	2299.02	2414.60	2601.86	2688.89	2783.65
	0.50	1432.03	1448.81	1473.68	1485.34	1496.31	2414.52	2514.17	2677.64	2750.76	2833.41
	0.75	1484.40	1495.75	1511.15	1518.45	1528.02	2531.32	2613.65	2749.67	2809.52	2876.01
	1.00	1540.71	1545.08	1556.88	1558.01	1560.74	2659.24	2715.92	2828.81	2867.60	2918.02
$C = 7$	0.00	1402.65	1435.57	1481.44	1503.27	1525.09	2360.65	2509.85	2754.12	2867.46	2989.80
	0.25	1455.01	1483.74	1522.14	1541.20	1559.75	2471.26	2608.39	2830.52	2933.96	3046.53
	0.50	1507.94	1529.61	1561.05	1575.91	1589.98	2586.77	2707.97	2906.33	2995.85	3096.29
	0.75	1560.33	1576.58	1598.55	1609.04	1621.71	2703.59	2807.47	2978.36	3054.62	3138.90
	1.00	1616.65	1625.91	1644.28	1648.61	1654.44	2831.52	2909.75	3057.50	3112.71	3180.93
$C = 8$	0.00	1478.52	1516.36	1568.78	1593.81	1618.73	2532.87	2703.61	2982.75	3112.51	3252.64
	0.25	1530.89	1564.53	1609.50	1631.77	1653.41	2643.50	2802.18	3059.18	3179.03	3309.41
	0.50	1583.85	1610.41	1648.42	1666.48	1683.65	2759.02	2901.77	3135.02	3240.94	3359.17
	0.75	1636.26	1657.41	1685.95	1699.63	1715.40	2875.86	3001.29	3207.05	3299.72	3401.79
	1.00	1692.59	1706.74	1731.68	1739.21	1748.14	3003.80	3103.58	3286.19	3357.82	3443.84
$C = 9$	0.00	1554.39	1597.15	1656.12	1684.35	1712.37	2705.09	2897.37	3211.38	3357.56	3515.48
	0.25	1606.77	1645.32	1696.86	1722.34	1747.07	2815.74	2995.97	3287.84	3424.10	3572.29
	0.50	1659.76	1691.21	1735.79	1757.05	1777.32	2931.27	3095.57	3363.71	3486.03	3622.05
	0.75	1712.19	1738.24	1773.35	1790.22	1809.09	3048.13	3195.11	3435.74	3544.82	3664.68
	1.00	1768.53	1787.57	1819.08	1829.81	1841.84	3176.08	3297.41	3514.88	3602.93	3706.75
$C = 10$	0.00	1630.26	1677.94	1743.46	1774.89	1806.01	2877.31	3091.13	3440.01	3602.61	3778.32
	0.25	1682.65	1726.11	1784.22	1812.91	1840.73	2987.98	3189.76	3516.50	3669.17	3835.17
	0.50	1735.67	1772.01	1823.16	1847.62	1870.99	3103.52	3289.37	3592.40	3731.12	3884.93
	0.75	1788.12	1819.07	1860.75	1880.81	1902.78	3220.40	3388.93	3664.43	3789.92	3927.57
	1.00	1844.47	1868.40	1906.48	1920.41	1935.54	3348.36	3491.24	3743.57	3848.04	3969.66

All values are expressed in dollars.

suggesting that those results are not sensitive to minor tweaks in workflow design.

Second, we undertook a three-server extension and tested several scenarios having 0, 1, and 2 of the servers being entirely sequestered ($\theta = 1$) across a few levels of α and λ_i. The results confirmed our expectations, and our finding above, that sequestering resources away from interruptions allows the system to provide better service to production work but degrades the system's service to interruptions. We

Figure 6 (a) Cost of Sequestering Relative to Cost of Not Sequestering (λ_i); (b) Cost of Sequestering Relative to Cost of Not Sequestering ($2\lambda_i$)

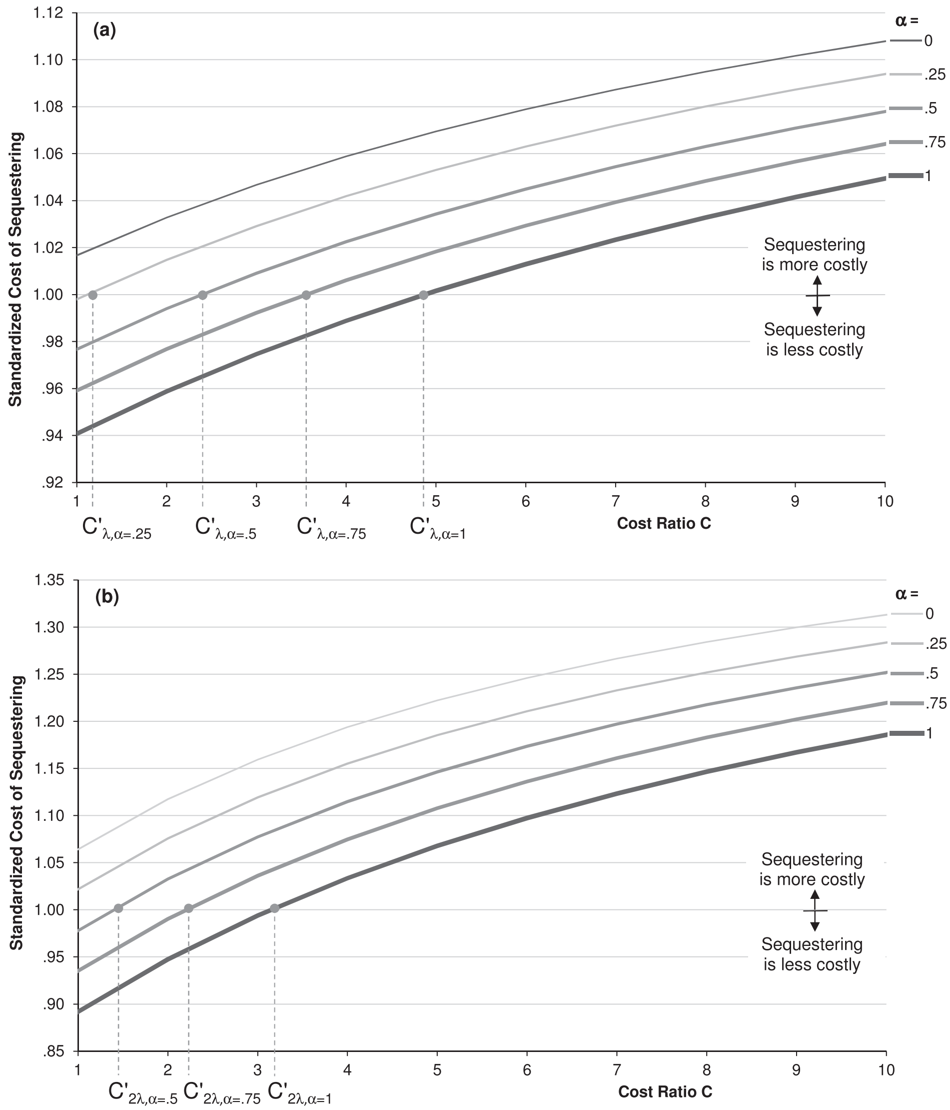
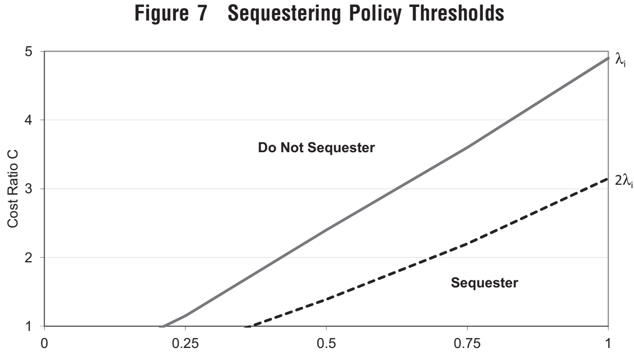


Figure 7 Sequestering Policy Thresholds



also observed that the change in these performance metrics between sequestering no servers and sequestering one server was larger than the change between sequestering one server and sequestering a second (additional) server. This is an intuitive result if we see sequestering as a shift of capacity; sequestering one server represents a larger *proportion* of additional capacity made unavailable to interruptions than does sequestering a second (or third, etc.) server. Applying the same cost modeling approach from our two-server analysis suggests that either very high forgetting rates or very low values of C (perhaps <1) would be required to justify sequestering two of three available servers, with this requirement growing as a larger proportion of servers is being considered for sequestering. In summary, none of the results suggested that our previous conclusions should be revised or qualified when thinking about how our findings apply beyond a two-server model. Neither space nor rigor requirements permit extensive discussion of the three-server results; given the many variations possible in this model (in terms of forgetting rates, variable sequestering levels, *different* forgetting rates, etc.), these initial, limited three-server sensitivity analyses indicate support for the generalizability of our two-server results.

5. Conclusions

5.1. Discussion

This study incorporated the concept of forgetting, which can produce re-work when interruptions disrupt routine production work performed by human labor. We explored the operational tactic of sequestering, or protecting from interruptions, a subset of servers in a knowledge-intensive service environment. The practice environment used to test the hypotheses was a radiology reading room. After gathering primary empirical data on arrivals and durations of exam readings and interruptions, we performed a series of simulations testing different levels of sequestering, re-work, and interruption arrival rate. We then augmented those simulation results with a cost ratio variable to examine sequestering effects on total system cost. Our findings indicate that the hypothesized relationships were generally supported, although some were conditionally dependent on re-work being explicitly accounted for in the model. In addition, we found a set of optimal sequestering policies for the scenarios under consideration. Taken collectively, those optimal policies can be stated as being dependent on values of C, the ratio of time costs for interruptions to those of production work, and α, the rate at which the servers forget; *lower levels of C and higher levels of α both increase the cost benefits of sequestering*.

In the scenarios where forgetting-induced re-work is not considered (i.e., $\alpha = 0$), our simulation results did not support Hypotheses 1 (production work flow time), 2 (production work wait time), and 5 (total shift duration). This is readily explained by considering key relationships in the model. Total flow time consists of waiting time and processing time. The production process in our study is such that, upon arrival, exams wait in a common queue for the next available resource. Sequestering guides a larger proportion of interruptions to the un-sequestered resource. However, both resources are still available for exams. When re-work is not considered, interruptions do not cause additional work above and beyond the sum of production work and interruptions, so sequestering does not alter the amount of total work to be done. In other words, when re-work is not considered, the design of the process (the operational policy of sequestering) has no influence on the total amount of effort (measured in time) required to complete all work. This is likely why we observed no significant change in flow time or waiting time for production work or total shift duration under the "no re-work" ($\alpha = 0$) scenarios.

While flow times and waiting times for routine production work were found to be reduced by sequestering, we also found that, as expected, flow times and waiting times for interruptions were increased significantly. By themselves, these results leave unanswered the question of whether sequestering can be considered a useful policy. Three additional findings suggest that it is. First, the total shift duration, as discussed above, was significantly reduced. This support for Hypothesis 5 provides confidence in the system-level benefits of the sequestering policy. However, we may still be uncomfortable with the fact that interruptions face significantly increased flow and/or wait times, for these jobs may require some immediacy.

Second, if we examine the *magnitude* of the increases in interruption flow and wait times created by sequestering (see Table 2), we find that the actual increases to interruption flow times (for λ_i) run consistently around 30 seconds (no sequestering vs. 100% sequestering), and the actual increases to interruption wait times (for those interruptions that were forced to wait) also look to be about 30 seconds or less. Therefore, the actual effect of sequestering on interruptions seems small when compared to the time savings provided to production work—up to a minute-and-a-half for both flow and wait times—suggesting that the trade-off may be a reasonable one, especially given that production work represents the large majority of the work being performed. These numbers increase when examining the $2\lambda_i$ condition—up to 75 seconds added to interruptions' flow time compared to 5 minutes shaved from exams—but only suggest an even more attractive trade-off (assuming, of course, that a 75-second delay to resolving an interruption is still reasonable). Indeed, in follow-up conversations with the radiology department about these findings, they agreed that such a small (in magnitude) increase in the time required to handle interruptions could often be justified by such a large (in comparison) time savings for exams. This is a rather interesting finding because it is the increase in interruptions that is making sequestering a resource away from those same interruptions more attractive. In other words, the faster the interruptions arrive, the bigger the overall time savings sequestering can offer.

Finally, when we incorporated cost information into the quantitative results, we found a more nuanced set of conclusions. Sequestering can indeed be a useful policy, but the decision to implement it is dependent on three factors: the cost ratio C, the forgetting rate α, and the interruption arrival rate λ_i. From Figure 7, we observe that if the environment we modeled has a cost ratio $C > 5$, sequestering should never be attempted, for the net result will be an increase in total costs. Moreover, if the workers exhibit forgetting rates at the low end of the α values we tested, then these critical cost ratios become irrelevant; sequestering is never desirable. If the environment becomes busier and interruptions arrive twice as quickly (i.e., $2\lambda_i$), the pattern of conclusions is similar, but the policy favors *not* sequestering at smaller values of C in all cases. This makes intuitive sense, as an increase in the number of interruptions should only magnify the effect of their time costs, thus reducing the value of C at which sequestering no longer provides overall cost reductions.

One interesting and important lesson for researchers can be taken from the fact that three of our hypotheses are supported only when forgetting-induced re-work is explicitly accounted for in the simulation. Had this model not specifically incorporated the cognitive work processes involved in processing these radiology exams and interruptions, we might have concluded quite erroneously that sequestering provides no operational or cost benefits. Instead, by modeling behavioral aspects of the system beyond those typically considered by process/workflow analyses, we have correctly discovered that a sequestering policy *can* provide reductions in overall work time and improvements to the execution of routine production work while not greatly hindering the system's ability to accommodate necessary interruptions. Moreover, we were able to characterize the optimal sequestering policies as dependent on both the cost ratios inherent to the system as well as the forgetting rates specific to the labor pool being investigated.

Overall, the operational benefits of sequestering in the radiology context studied appear to be clear: sequestering has the potential to reduce overall production work flow and wait times while degrading only slightly the service level provided to interruptions. However, the decision to sequester must take into account the ratio of costs associated with delaying interruptions vs. the costs associated with delaying production work; when this ratio becomes high and/or when forgetting rates are extremely low, sequestering becomes less desirable. This research also demonstrated the significant value of considering the operational effect of interruptions above and beyond any psychological detriment they may pose, such as frustration (Perlow 1998). Routine production work combined with handling interruptions can result in more *total* effort than just the sum of both activities due to the forgetting/re-learning/re-work that occurs in knowledge-intensive service environments that are subjected to frequent interruption.

However, these findings should be viewed as early support for the importance of considering forgetting-induced re-work in service operations research; while we have started to fill a gap in the knowledge-work literature, additional research will be needed to fully understand these phenomena. Industries as diverse as professional services (e.g., health care, legal, financial, and consulting), educational services, complex services like airlines and commercial construction management, and even public administration, security, and defense services could easily be subject to these interruption effects due to their reliance on knowledge-intensive work.

5.2. Managerial Implications

Managerially, these results suggest some interesting prescriptions. First, when comparing the λ_i and $2\lambda_i$ policy curves in Figure 7, as interruptions arrive more quickly ($2\lambda_i$), sequestering becomes less desirable. However, it is precisely when interruptions are coming "fast and furious" that employees will be most tempted to sequester themselves (i.e., close their doors and hide). Our results suggest that this is the *opposite* of desired behavior in situations where delaying interruptions is relatively costly (high values of C). If management can monitor the work environment and be sensitive to heightened levels of interruptions and then take appropriate action, that should help the organization's knowledge workers maintain their productivity, if not also reduce stress and sustain high levels of job satisfaction.

In addition, these results suggest that service organizations would be wise to consider sequestering first those workers who have higher forgetting rates and work in environments with low values of cost ratio C. This implies that organizations may wish to start undertaking two practices that are not, to our knowledge, currently common. First, employees who may be subject to recurrent interruptions should be evaluated for their ability to retain details under these circumstances. Second, the organization may wish to study its work to determine how costly it is to delay routine production work as well as to delay incoming interruptions. One approach could be the totaling of the time for all downstream resources that could be rendered idle by virtue of a delay in the completion of this job, but other costs that are more difficult to quantify (e.g., customer satisfaction) may also influence these estimates. Combined, these values provide an estimate for the work environment's cost ratio C, which is useful for identifying the preferred sequestering policy. Empirically estimating a worker's α (forgetting rate) and the organization's cost ratio C would be helpful in ensuring that work systems and policies are as aligned as possible with the characteristics of the employees involved. In situations where sequestering is beneficial and the employees have different forgetting rates, *ceteris paribus*, the organization would benefit more by sequestering the employee with the larger (or largest) value of α.

5.3. Limitations and Extensions

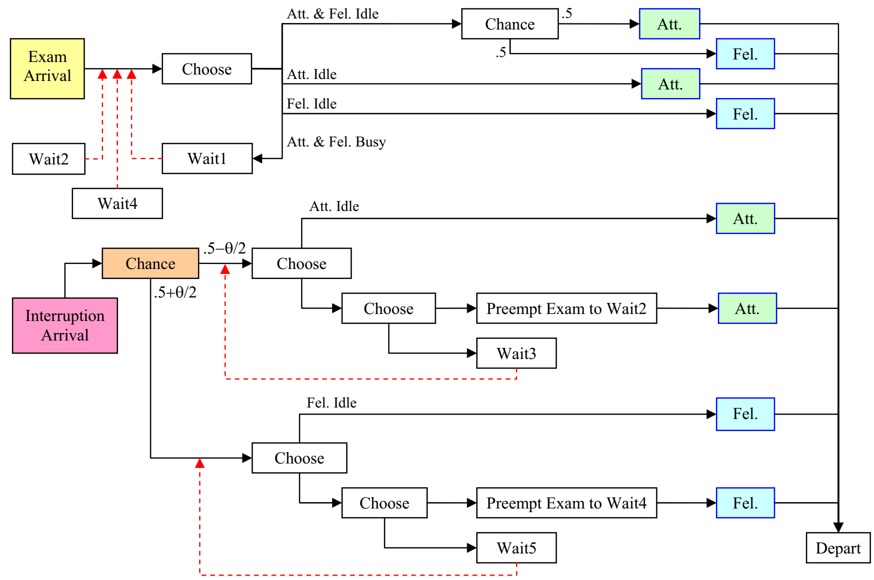
As with any study, this research has limitations and could be extended in some important ways. First, our conclusions regarding sequestering and its effects due to forgetting and re-work are based on modeling one environment in a single organization, which limits the study's generalizability. A more comprehensive study involving a larger sample and/or models of work systems from a variety of service settings would provide more generalizable insights. However, despite modeling just one service, these results seem to suggest implications for service process design (Hill et al. 2002) on a variety of issues. For example, as knowledge-intensive work is often necessary in highly customized services, the use of scripting (a specification of actions the employee must perform) might help recovery from interruptions by accelerating the re-acquisition of forgotten task knowledge. However, scripting has been shown to negatively affect perceived service quality in highly customized services (Victorino et al. 2013), so scripting may not be the best mechanism to counteract interruption-induced forgetting. An alternative to scripting may be to impose a structure as to how and when interruptions can be accommodated. Consistent with the literature on service sequencing (e.g., Dixon and Verma 2013), which has yielded substantial evidence that certain service sequences provide more utility to customers than others, there may be a policy that reduces interruptions while also improving the delayed customer's perception of the overall service experience.

Second, the resources considered here are assumed to be homogeneous (i.e., processing of exams and interruptions can be done by either of the resources with equal efficiency and forgetting rates were identical) and constant. This is unlikely to be strictly true in many knowledge-intensive service organizations, as the ability of knowledge workers can vary greatly from individual to individual and even from moment to moment (Xia and Sudharshan 2002), depending on a variety of factors (e.g., experience, physical health/state, mental health/state, work environment, etc.). This diversity makes these factors difficult to include experimentally, so this first foray into studying the operational effects of sequestering considered only homogeneous resources. While we attempted to mitigate this limitation by considering a wide range of values for α (from 0 to 1), this issue certainly represents significant opportunity for future researchers to find ways to empirically measure α in a variety of knowledge workers and work environments so as to better understand what range of our results is likely to be most relevant to practice.

Third, in our study, an interruption cannot pre-empt another interruption (interruptions have no priority over one another beyond FCFS). This could be viewed as a limitation. However, because interruptions themselves are not subject to forgetting effects, this assumption ensures that our findings reflect the most *conservative* (smallest) estimate of the effects of within-job forgetting on overall system performance. If interruptions were allowed to be pre-empted by other interruptions, this would inflate the forgetting-induced re-work and degrade system performance even further. As there is some evidence that knowledge workers can face multiple, even tiered levels of interruptions (Tucker and Spear 2006), estimating the additional performance degradation this phenomenon would generate appears to be one of several opportunities for future extensions of this research.

Finally, while this study did not specifically consider the effect of interruptions on the *quality* of the work performed—just the quantity—there is increasing evidence that even short interruptions increase the chance of a worker making an error (Altmann et al. 2013). Therefore, it is possible that the total effect of interruption-induced forgetting could be to both delay the completion of the task and to create an error, which might then cause additional re-work and even further delay. This suggests that future work in this area may benefit from measuring the *quality-weighted* productivity of interrupted knowledge workers.

Appendix: Logic Diagram for Simulation Model



Note. Att. = Attending; Fel. = Fellow; θ = sequestering level; *Wait1* = common exam queue; *Wait2* = preempted exam queue for Att.; *Wait3* = interruption queue for Att.; *Wait4* = preempted exam queue for Fel.; *Wait5* = interruption queue for Fel.

Notes

[1]While there exist many definitions of and metrics for "productivity" in the context of knowledge workers (see Marucheck and Sulek 1993 and Ramirez and Nembhard 2004 for excellent reviews), our use here is limited to the quantity of work completed per time period with or without considering the cost of resources involved.

[2]We assume $C_i \geq C_x$ because a rational (cost-minimizing) worker would not delay regular production work by attending to an interruption if the interruption's importance (i.e., time-unit cost of delaying the interruption) was lower than that of the production work. Such low-priority interruptions would likely be attended to at the conclusion of the job at hand, thereby negating the forgetting penalty and rendering them analytically uninteresting.

[3]While some lines in Table 3 have values at a midpoint lower than both endpoints, inspection shows that these values are always within the confidence interval of the less costly endpoint, making them statistically equivalent to the optimal policy for that combination of C, α, and λ_i.

References

Allwood, J. M., W. L. Lee. 2004. The impact of job rotation on problem-solving skills. *Int. J. Prod. Res.* **42**(5): 865–881.

Altmann, E. M., J. G. Trafton, D. Z. Hambrick. 2013. Momentary interruptions can derail the train of thought. *J. Exp. Psychol. Gen.* (in press).

Apte, U., R. Cavalier, S. Kulkarni. 2010. Analysis and improvement of information intensive services: Empirical evidence from insurance claims handling operations. *Prod. Oper. Manag.* **18**(1): 665–678.

Ash, R., D. E. Smith-Daniels. 1999. The effects of learning, forgetting, and relearning on decision rule performance in multiproject scheduling. *Decision Sci.* **30**(1): 47–82.

Bailey, C. 1989. Forgetting and the learning curve: A laboratory study. *Manage. Sci.* **35**(3): 340–352.

Bailey, C., E. V. McIntyre. 2003. Using parameter prediction models to forecast post-interruption learning. *IIE Trans.* **35**(12): 1077–1090.

Bannister, F., D. Remenyi. 2009. Multitasking: The uncertain impact of technology on knowledge workers and managers. *Electron. J. Inf. Syst. Eval.* **12**(1): 1–12.

Benkard, C. L. 2000. Learning and forgetting: The dynamics of aircraft production. *Am. Econ. Rev.* **90**(4): 1034–1054.

Bitner, M. J., S. W. Brown. 2008. The service imperative. *Bus. Horiz.* **51**(1): 39–46.

Bitner, M. J., S. W. Brown, M. L. Meuter. 2000. Technology infusion in service encounters. *J. Acad. Mark. Sci.* **28**(1) 138–149.

Brockner, J., B. Wiesenfeld. 1996. An integrative framework for explaining reactions to decisions: Interactive effects of outcomes and procedures. *Psychol. Bull.* **120**(2) 189–208.

Carlson, J. G., A. J. Rowe. 1976. How much does forgetting cost? *Ind. Eng.* **8**(9): 40–47.

Chase, R. 1981. The customer contact approach to services: Theoretical bases and practical extensions. *Oper. Res.* **29**(4): 698–706.

Chiu, H. N., H. M. Chen, L. C. Weng. 2003. Deterministic time-varying demand lot-sizing models with learning and forgetting in setups and production. *Prod. Oper. Manag.* **12**(1): 120–127.

Cook, L. S., D. E. Bowen, R. B. Chase, S. Dasu, D. M. Stewart, D. A. Tansik. 2002. Human issues in service design. *J. Oper. Manag.* **20**(2): 159–174.

Dar-El, E. M., K. Ayas, I. Gilad. 1995. Predicting performance times for long cycle time tasks. *IIE Trans.* **27**(3): 272–281.

Davenport, T. H., R. J. Thomas, S. Cantrell. 2002. The mysterious art and science of knowledge-worker performance. *Sloan Manag. Rev.* **44**(1): 23–30.

Davis, G. B., J. D. Naumann. 1999. Knowledge work productivity. K. Kendall, ed. *Emerging Information Technologies: Improving*

Decisions, Cooperation, and Infrastructure. Sage, Thousand Oaks, CA, 343–357.

Ding, D. X., P. J.-H. Hu, R. Verma, D. G. Wardell. 2010. The impact of service system design and flow experience on customer satisfaction in online financial services. *J. Serv. Res.* **13**(1): 96–110.

Dixon, M., R. Verma. 2013. Sequence effects in service bundles: Implications for service design and scheduling. *J. Oper. Manage.* **31**(3): 138–152.

Drucker, P. F. 1991. The new productivity challenge. *Harv. Bus. Rev.* **69**(6): 69–79.

Federgruen, A., L. Green. 1986. Queueing systems with service interruptions. *Oper. Res.* **34**(5): 752–768.

Finkenbinder, E. O. 1913. The curve of forgetting. *Am. J. Psychol.* **24**(1): 8–32.

Flynn, E. A., K. N. Barker, J. T. Gibson, R. E. Pearson, B. A. Berger, L. A. Smith. 1999. Impact of interruptions and distractions on dispensing errors in an ambulatory care pharmacy. *Am. J. Health Syst. Pharm.* **56**(13): 1319–1325.

Froehle, C. M., A. V. Roth. 2004. New measurement scales for evaluating perceptions of the technology-mediated customer service experience. *J. Oper. Manag.* **22**(1): 1–22.

Gaver, Jr., D. P. 1962. A waiting line with interrupted service, including priorities. *J. Roy. Stat. Soc. B* **24**: 73–90.

Globerson, S., N. Levin, A. Shtub. 1989. The impact of breaks on forgetting when performing a repetitive task. *IIE Trans.* **21**(4): 376–381.

Groenevelt, H., L. Pintelon, A. Seidmann. 1992. Production lot sizing with machine breakdowns. *Manage. Sci.* **38**(1): 104–123.

Halsted, M. J., C. M. Froehle. 2008. Design, implementation, and assessment of a radiology workflow management system. *Am. J. Roentgenol.* **191**(2): 321–327.

Healey, A. N., N. Sevdalis, C. A. Vincent. 2006. Measuring intra-operative interference from distraction and interruption observed in the operating theatre. *Ergonomics* **49**(5–6): 589–604.

Hill, A. V., D. A. Collier, C. M. Froehle, J. C. Goodale, R. D. Metters, R. Verma. 2002. Research opportunities in service process design. *J. Oper. Manage.* **20**(2): 189–202.

Hopp, W. J., S. M. R. Iravani, F. Liu. 2009. White collar workforce management: An operations-oriented survey. *Prod. Oper. Manag.* **18**(1): 1–32.

Hyer, N. L., U. Wemmerlöv, J. A. Morris. 2009. Performance analysis of a focused hospital unit: The case of an integrated trauma center. *J. Oper. Manag.* **27**(3): 203–219.

Jaber, M. Y., M. Bonney. 1997. A comparative study of learning curves with forgetting. *Appl. Math. Model.* **21**: 523–531.

Jaber, M. Y., M. Bonney. 2003. Lot sizing with learning and forgetting in set-ups and in product quality. *Int. J. Prod. Econ.* **83**(1): 95–111.

Jaber, M. Y., H. V. Kher, D. J. Davis. 2003. Countering forgetting through training and deployment. *Int. J. Prod. Econ.* **85**: 33–46.

Jett, Q., J. George. 2003. Work interrupted: A closer look at the role of interruptions in organizational life. *Acad. Manag. Rev.* **28**(3): 494–507.

Kelton, W. D., R. P. Sadowski, D. T. Sturrock. 2004. *Simulation with Arena*. McGraw-Hill, New York, NY.

Kher, H. V., M. K. Malhotra, P. R. Philipoom, T. D. Fry. 1999. Modeling simultaneous worker learning and forgetting in dual resource constrained systems. *Eur. J. Oper. Res.* **115**: 158–172.

Marucheck, A., J. Sulek. 1993. A process approach to measuring knowledge worker productivity. *Oper. Manage. Rev.* **9**(4): 1–15.

Matson, E., L. Prusak. 2010. Boosting productivity of knowledge workers. *McKinsey Q.* **4**: 93–96.

Metters, R., V. Vargas. 2000. A typology of de-coupling strategies in mixed services. *J. Oper. Manag.* **18**(6): 663–682.

Metters, R., R. Verma. 2008. History of offshoring knowledge services. *J. Oper. Manag.* **26**(2): 141–147.

Meuter, M. L., A. L. Ostrom, R. I. Roundtree, M. J. Bitner. 2000. Self-service technologies: Understanding customer satisfaction with technology-based service encounters. *J. Mark.* **64**(3): 50–64.

Moinzadeh, K., P. Aggarwal. 1997. Analysis of a production/inventory system subject to random disruptions. *Manage. Sci.* **43**(11): 1577–1588.

Özbayrak, M., G. Cagil, C. Kubat. 2004. How successfully does JIT handle machine breakdowns in an automated manufacturing system? *J. Manuf. Technol. Manag.* **15**(6): 479–494.

Pang, G., W. Whitt. 2009. Service interruptions in large-scale service systems. *Manage. Sci.* **55**(9): 1499–1512.

Perlow, L. A. 1998. Finding time, stopping the frenzy. *Bus. Health* **16**(18): 31–35.

Perlow, L. A. 1999. The time famine: Toward a sociology of work time. *Adm. Sci. Q.* **44**(1): 57–81.

Piccoli, G., M. K. Brohman, R. T. Watson, A. Parasuraman. 2009. Process completeness: Strategies for aligning service systems with customers' service needs. *Bus. Horiz.* **52**(4): 367–376.

Raheja, A. S., V. Subramaniam. 2002. Reactive recovery of job shop schedules: A review. *Int. J. Adv. Manuf. Technol.* **19**: 756–763.

Ramirez, Y. W., D. A. Nembhard. 2004. Measuring knowledge worker productivity: A taxonomy. *J. Intellectual Capital* **5**(4): 602–628.

Rao, S. S. 1965. Queuing models with balking, reneging, and interruptions. *Oper. Res.* **13**(4): 596–608.

Sampson, S. E., C. M. Froehle. 2006. Foundations and implications of a proposed unified services theory. *Prod. Oper. Manag.* **15**(2): 5–38.

Schmenner, R. W., M. L. Swink. 1998. On theory in operations management. *J. Oper. Manag.* **17**(1): 97–113.

Seshadri, S., Z. Shapira. 2001. Managerial allocation of time and effort: The effects of interruptions. *Manage. Sci.* **45**(5): 647–662.

Sikström, S. 2002. Forgetting curves: Implications for connectionist models. *Cogn. Psychol.* **45**: 95–152.

Skinner, B. F. 1974. *About Behavioralism*. Knopf, New York, NY.

Smunt, T. L., J. Meredith. 2000. A comparison of direct cost savings between flexible automation and labor with learning. *Prod. Oper. Manag.* **9**(2): 158–170.

Sparks, B., J. McColl-Kennedy. 1998. The application of procedural justice principles to service recovery attempts: Outcomes for customer satisfaction. *Adv. Consum. Res.* **25**(1): 156–161.

Speier, C., I. Vessey, J. S. Valacich. 2003. The effects of interruptions, task complexity, and information presentation on computer-supported decision-making performance. *Decis. Sci.* **34**(4): 771–797.

Teyarachakul, S., S. Chand, J. Ward. 2011. Effect of learning and forgetting on batch sizes. *Prod. Oper. Manag.* **20**(1): 116–128. doi: 10.1111/j.1937-5956.2010.01140.x.

Tucker, A. L., S. J. Spear. 2006. Operational failures and interruptions in hospital nursing. *Health Serv. Res.* **41**(3): 643–662.

Victorino, L., R. Verma, D. G. Wardell. 2013. Script usage in standardized and customized service encounters: Implications for perceived service quality. *Prod. Oper. Manag.* **22**(3): 518–534.

White, H. C., L. S. Christie. 1958. Queueing with preemptive priorities or with breakdown. *Oper. Res.* **6**(1): 79–95.

Woltz, D. J., V. J. Shute. 1995. Time course of forgetting exhibited in repetition priming of semantic comparisons. *Am. J. Psychol.* **108**(4): 499–525.

Xia, L., D. Sudharshan. 2002. Effects of interruptions on consumer online decision processes. *J. Consumer Psychol.* **12**(3): 265–280.

Yelle, L. E. 1979. The learning curve: Historical review and comprehensive survey. *Decis. Sci.* **10**(2): 122–128.

Zellmer-Bruhn, M. E. 2003. Interruptive events and team knowledge acquisition. *Manage. Sci.* **49**(4): 514–528.