Arab Journal of Administrative Sciences

2023, Vol. 30, No. 3, 555-584
Doi: 10.34120/ajas.v30i3.233

Ahmad M. Ashkanani
Kuwait University
Kuwait

The Productivity Spectrum: An Individual-Level Examination of Servers' Reactions to Workload and Overwork

Abstract

Purpose: This study explores the diverse reactions of human servers to variations in workload and overwork, focusing on how these stressors influence service time in a call center setting.

Study design/methodology/approach: OLS regressions with fixed effects and clustered standard errors were used to analyze the impact of workload and overwork on service time at both aggregate and individual levels.

Sample and data: This research utilizes archival data from a call center at a large Kuwaiti private hospital, covering 58,460 calls handled by 23 agents during October 2016.

Results: At the aggregate level, workload has a negative effect on service time, indicating a general speedup response, while overwork shows no significant effect. At the individual level, responses vary significantly among agents, with some agents speeding up, others slowing down, and some demonstrating non-linear responses to workload. Overwork responses also vary, with some agents slowing down, others speeding up, and some showing flat responses.

Originality/value: This study contributes to the behavioral queuing literature by emphasizing the importance of individual-level analysis in understanding server responses to workload and overwork. The findings reveal that aggregate analyses can mask significant variations in individual behaviors, crucial for optimizing call center performance. By exposing the spectrum of individual reactions, the study provides deeper theoretical and practical insights into the complex dynamics of service time under different work stressors.

Research limitations/implications: The findings provide insights for optimizing call center performance, but further research is needed to validate and generalize these findings across different settings and call center environments.

Keywords: Productivity, Workload, Overwork, Variability, Individual-Level Analysis, Call Center Operations.

JEL classification: L23, M19, D9, M54, J24

Submitted: 31/12/2023, revised: 17/6/2024, accepted: 18/7/2024.

Published by the Academic Publication Council of Kuwait University. All rights reserved.

To cite: Ashkanani, A. M. (2023). The productivity spectrum: An individual-level examination of servers' reactions to workload and overwork. *Arab Journal of Administrative Sciences*, *30*(3), 555-584. https://doi.org/10.34120/ajas.v30i3.233

الملخص

طيف الإنتاجية: دراسة تحليلية لتفاعل الأفراد مع أعباء العمل والإجهاد

أحمد محمد أشكناني

جامعة الكويت، الكويت

هدف الدراسة: تهدف الدراسة إلى استكشاف التباين في تفاعل موظفي خدمة العملاء مع أعباء العمل والإجهاد، وكيفية تأثير هذه الضغوطات على الوقت المستغرق في خدمة عملاء مراكز الاتصال.

تصميم/ منهجية/ طريقة الدراسة: تستخدم هذه الدراسة نماذج تجريبية، تُحلل باستخدام سلسلة من نماذج الانحدار ذات التأثيرات الثابتة (Fixed Effects)، والأخطاء المعيارية المجمعة (Clustered Stand-ard Errors) لدراسة تأثير أعباء العمل والإجهاد على المدة الزمنية لخدمة العملاء عن طريق التحليلات الكلية والفردية.

عينة الدراسة وبياناتها: تستند هذه الدراسة إلى بيانات أرشيفية من مركز اتصال في مستشفى خاص بدولة الكويت، حيث حُللت بيانات 58460 مكالمة تم التعامل معها بوساطة 23 موظف خدمة عملاء خلال شهر أكتوبر من عام 2016.

نتائج الدراسة: تكشف نتائج الدراسة عن تباين كبير في كيفية استجابة موظفي خدمة العملاء لأعباء العمل والإجهاد، فقد أظهرت النتائج أن بعض الموظفين يستجيبون لزيادة ضغوط العمل بتسريع وتيرته، بينما يبطئ البعض الآخر وتيرة العمل أو يستجيبون بطريقة غير خطية للتغيرات في أعباء العمل والإجهاد على رغم من تشابه ظروفه.

أصالة الدراسة: تسهم هذه الدراسة في إثراء أدبيات سلوكيات طوابير الانتظار من خلال تبيان أهمية تحليل السلوكيات على مستوى الفرد لفهم التباين في استجابة الموظفين لأعباء العمل والإجهاد، حيث كشفت النتائج أن التحليلات الكلية يمكن أن تخفي اختلافات جوهرية في سلوكيات موظفي خدمة العملاء، مما يسلط الضوء على أهمية الفهم العميق للديناميكيات المعقدة للسلوكيات الفردية لموظفي خدمة العملاء واستجابتهم لضغوط العمل المختلفة.

حدود الدراسة وتطبيقاتها: توفر نتائج الدراسة رؤى للمديرين لتحسين الأداء التشغيلي لمراكز الاتصال، وتؤكد الحاجة لمزيد من الأبحاث؛ للتحقق من هذه النتائج وتعميمها عبر بيئات الأعمال المختلفة.

الكلمات المفتاحية: الإنتاجية، أعباء العمل، الإجهاد، التباين، التحليلات الفردية، عمليات مراكز الاتصال.

تصدر عن مجلس النشر العلمي بجامعة الكويت. جميع الحقوق محفوظة للمجلة.

الإشارة المرجعية: أشكناني، أحمد محمد. (2023). طيف الإنتاجية: دراسة تحليلية لتفاعل الأفراد مع أعباء العمل والإجهاد. *المجلة العربية للعلوم الإدارية*، *30*(3)، 555-584.

https://doi.org/10.34120/ajas.v30i3.233

Introduction

Understanding server behavior in queuing systems is critical for improving the performance of service systems. Research suggests that servers' behavior impacts the performance of queuing systems (Allon & Kremer, 2018). In particular, servers' strategic behavior in response to system-level factors such as queue structure (Wang & Zhou, 2018), financial incentives (Shunko et al., 2017), and congestion levels (Kc & Terwiesch, 2009) have been the focus of research in the recent years. Indeed, the operations management field is increasingly concentrating on studying the behavioral dynamics of individual workers rather than using simplistic models that view workers as a uniform group of "automatons" with "simple and fairly predictable work patterns" (Bendoly & Prietula, 2008, p. 1131). Several studies in behavioral queuing literature have investigated servers' reactions to work stressors, such as workload and overwork, also known as instantaneous and extended load (Delasay et al., 2019), and their impact on server productivity. For example, Kc and Terwiesch (2009) found that healthcare service workers speed up as system workload increases, but this acceleration is unsustainable and can lead to overwork, which lowers service rates and increases mortality.

Building on this foundational knowledge, recent studies have begun to explore more nuanced aspects of server behavior. However, there have been mixed findings in the literature regarding the direction of the workload-productivity relationship. Evidence suggests that the relationship is more nuanced than originally assumed (Delasay et al., 2019), and multiple factors could influence the direction and shape of the workload-productivity relationship, including financial incentives (Tan & Netessine, 2014) and operational context (Berry Jaeker & Tucker, 2017). For example, Tan and Netessine (2014) demonstrated that restaurant chain servers balance sales and speed efforts based on workload, and when workload increases beyond a threshold, servers accelerate and reduce their sales efforts (potentially at the expense of receiving lower tips) to avoid high waiting costs perceived by customers. This implies that the relationship between workload and service time (an inverse proxy for productivity) exhibits an inverted U-shape and is influenced by servers' internalized incentives. In a healthcare operational context, Berry Jaeker and Tucker (2017) found that patient service time increases as workload increases until a tipping point, "after which patients are discharged early to alleviate congestion" (Berry Jaeker & Tucker, 2017, p. 1042), and a second tipping point, where additional occupancy leads to a longer service time. These

results suggest an N-shaped relationship between workload and service time and that hospitals may speed up or slow down patients' service time depending on both the system state (i.e., congestion level) and the needs of patients (e.g., high vs. low severity patients). These varied findings underscore the complexity of human responses to workload and the need for a deeper understanding of individual-level behaviors.

Despite the wealth of research on workload and productivity, significant knowledge gaps remain. Most studies focus on aggregate-level analyses of servers' responses to workload, with few examining the effects of overwork in the same detail. While this approach is useful for making generalizations about servers' responses to work stressors, it may mask important variations in the distribution of these responses. For example, while studies in the behavioral queuing literature analyze large datasets with thousands (or hundreds of thousands) of transactions, these transactions are typically handled by a smaller number of human workers, ranging between approximately 20-100 workers per study[1] (see Table 1). This, in turn, might lead to biases in the estimates of work stressor effects on servers' productivity, as these aggregate-level estimates are generated from a sample of servers that do not necessarily represent the overall behavior of the population of human servers. Consequently, this may contribute to conflicting results in the literature regarding the direction and magnitude of the relationship between service time and work stressors.

This study aims to address these gaps by proposing and testing an individual-level approach to examine the effects of workload and overwork on human servers' behavior. The central research question is: How do individual servers in a call center setting react to workload and overwork, and to what extent do the effects of workload and overwork on service time vary across service agents? To investigate this, a set of competing hypotheses on the impact of workload on service time is formulated, informed by insights from the behavioral queuing literature. Utilizing econometric models with fixed effects and clustered standard errors, these hypotheses are tested by analyzing archival data from a call center at a large Kuwaiti private hospital, comprising over 58,000 calls handled by 23 agents. The study begins with a baseline aggregate-level examination of the workload effect, which is then contrasted with the proposed individual-level approach. Furthermore, considering the limited research indicating that overwork generally

[1] Some studies even fail to report the number of human servers included in their sample.

leads to slower service times, this hypothesis is similarly investigated. This research contributes to the behavioral queuing literature by offering a more detailed examination of the distribution of individual responses to work stressors. This approach has the potential to reduce biases in the estimates of work stressor effects on human server behavior and may help explain some of the discrepancies found in the literature. Additionally, it provides practical insights for optimizing call center performance through tailored management strategies, as discussed in later sections of the paper.

Literature Review and Hypotheses Development

The response of human servers to variations in workload and overwork is a complex phenomenon influenced by numerous factors. The behavioral queuing literature provides a diverse array of findings regarding how workload impacts service time, revealing relationships characterized by *speedup* (Ashkanani et al., 2022; Freeman et al., 2017; Kc & Terwiesch, 2009, 2012; Kc et al., 2020; Staats & Gino, 2012), *slowdown* (Ashkanani et al., 2022; Batt & Terwiesch, 2016; Kc, 2013), *non-linear* (Ashkanani et al., 2022; Batt & Terwiesch, 2016; Berry Jaeker & Tucker, 2017; Tan & Netessine, 2014; Wang & Zhou, 2018), and even *flat* (Freeman et al., 2017; Song et al., 2015) responses to workload. Some studies report different effects within the same study (e.g., depending on the service encounter segments), illustrating the nuanced impact of workload conditions. As for overwork, fewer studies have examined its effects, but those that do generally agree on a slowdown effect as work demands become excessive (Kc & Terwiesch, 2009; Staats & Gino, 2012).

This section delves deeper into the behavioral queuing literature by contrasting the various perspectives and formulating competing hypotheses based on the diverse findings. Notably, while the literature mentions flat responses to workload, this study will focus on patterns where significant changes in service time are evident—namely, *speedup*, *slowdown*, and *non-linear* effects. This approach aims to concentrate on the dynamics where operational adjustments and managerial interventions might be most impactful. Table 1 summarizes the results of these studies, providing a comparative overview to further facilitate the analysis.

Table 1
Summary of Workload and Overwork Effects on Service Time from the Behavioral Queuing Literature

Setting	Source	Sample Size		Task Description	Work Stressor Effect on Service Time	
		No. of Servers	No. of Transactions		Workload	Overwork
Health-care services	Kc and Terwiesch (2009)	26 transporters (study 1); N/A (study 2)	17,000 patient transports (study 1); 2,740 patients (study 2)	Physical transportation of patients within hospital premises (study 1); Cardiothoracic surgeries involving multiple care providers (study 2)	Speedup (Linear)	Slowdown (Linear)
	Kc and Terwiesch (2012)	N/A	1,036 patient admissions	Cardiothoracic ICU stays involving multiple care providers	Speedup (Linear)	-
	Song et al. (2015)	40 ED physicians	217,161 ED patient visits	Treatment of ED patients	Flat	-
	Batt and Terwiesch (2016)	N/A	12,260 patient visits	Treatment of ED "abdominal pain" patients	Slowdown (Linear) & Inverted U-shape based on workload definition	-
	Kc (2013)	26 physicians	145,935 ED patient visits	Treatment of ED patients	Slowdown (Linear)	-
	Berry Jaeker and Tucker (2017)	N/A	241,198 patient visits	Treatment of hospital patients	N-shape	-
	Kc et al. (2020)	84 physicians	137,676 ED patient visits	Treatment of ED patients	Speedup (Linear) & Slowdown (Linear) based on workload definition	-

Cont. Table 1
Summary of Workload and Overwork Effects on Service Time from the Behavioral Queuing Literature

Setting	Source	Sample Size		Task Description	Work Stressor Effect on Service Time	
		No. of Servers	No. of Transactions		Workload	Overwork
	Freeman et al. (2017)	N/A	up to 16,355 births	Maternity delivery services	Speedup (Linear) for non-complex cases; Flat for complex cases	-
Financial services	Staats and Gino (2012)	111 bank workers	598,393 loan-related transactions	Processing loan-related transactions	Speedup (Linear)	Slowdown (Linear)
Restaurant services	Tan and Netessine (2014)	N/A	190,799 check-level transactions	Serving restaurant customers	Inverted U-shape	-
Call center services	Ashkanani et al. (2022)	82 call center agents	109,796 calls	Processing call center service requests	Inverted U-shape for offline ST, Speedup (Linear) for online ST, Slowdown (Linear) for total ST	Speedup* (Linear) for offline ST, Slowdown (Linear) for online ST, Flat* for total ST
Supermarket services	Wang and Zhou (2018)	N/A	3,245 checkout transactions	Checking out supermarket customers	Convex decreasing	-

Notes: N/A signifies that the number of servers was unavailable in the study. ST stands for service time. * The baseline model results of Ashkanani et al. (2022) indicated a slowdown response to overwork; however, the instrumental variables models indicated a speedup and a flat response to overwork for offline and total service times, respectively.

Workload: Speedup

Empirical evidence suggests that servers increase their work rate under high workloads, thereby reducing service time. For instance, in a study of two distinct healthcare services, Kc and Terwiesch (2009) examined the effect of workload on service time. First, they monitored 26 transporters responsible for moving patients within hospital facilities, analyzing a total of 17,000 patient transports. Their findings indicated that a 20% increase in transporter workload led to a 30-second reduction in transport time. Additionally, in a study of 2,740 cardiothoracic surgeries involving various care providers, the authors found that a 10% increase in workload resulted in a reduction of hospital length of stay by two days. While not directly tested, the authors suggested *rushing* and *task reduction* as potential mechanisms that explain the behavior of servers under increased workload. In particular, as the workload on the system increases, transporters are likely to experience longer waiting times. The authors suggest that speeding up the transport process helps in mitigating the longer waiting times. Consequently, transporters, whose productivity is continually monitored via an electronic system, are motivated to increase their speed when workload rises. In the case of cardiothoracic surgery, the authors proposed that under high workload conditions, hospitals might discharge patients earlier than usual to free up capacity for new patients. While this speeds up the discharge process, it may also lead to premature discharges that could negatively affect patient safety.

In another study, Kc and Terwiesch (2012) examined how ICU occupancy levels influence patient discharge decisions. They found that higher ICU occupancy leads to an aggressive discharge strategy, resulting in patients being discharged 16% earlier than those under lower occupancy conditions. The authors' findings confirm that staff members expedite patient flow when the ICU nears full capacity, demonstrating a direct relationship between workload and speedup behavior in healthcare service environments.

Similarly, Staats and Gino (2012) examined the behavior of 111 bank workers who processed around 600,000 loan-related transactions at a Japanese bank. While the study primarily investigated the interplay between specialization and variety in repetitive tasks, workload was controlled as a variable and showed significant effects on productivity. In particular, the results suggest that a 1 standard deviation increase in workload led to an 8.3% reduction in service time.

In a similar vein, Freeman et al. (2017) investigated the impact of workload on midwives in a maternity unit. They found that as the workload increased, midwives rationed discretionary services like epidural analgesia for noncomplex cases, reducing epidural provision by 28.8%. For complex cases, midwives referred more patients to specialists for physician-led deliveries, increasing referrals by 14.2%. While service time for complex cases remained consistent irrespective of workload levels, it decreased by 8.3% for noncomplex cases as the workload increased from low to high levels due to *task reduction* (i.e., rationing epidural administration).

Moreover, Kc et al. (2020) have studied the impact of three types of workload on the productivity of 84 physicians in an emergency department, namely: physician workload (quantified by the number of patients a particular physician is responsible for at any given moment), system load (quantified by the total workload being managed by physicians), and wait load (quantified by the number of patients awaiting pickup by a physician). The authors found that higher levels of physician and wait loads were associated with higher productivity in the short-term, while system load had the opposite effect. Notably, workload influenced physician's task completion preference, where higher levels of physician load increased the likelihood of physicians picking up easy cases, while both wait and system load had the opposite effect.

Finally, a recent study by Ashkanani et al. (2022) explored the effects of workload on the performance of 82 call center agents across approximately 100,000 service calls. The research segmented service time into three categories: online (active service time where the customer is connected to the agent on the line), offline (post-call processing time performed by the agent to wrap up the service request), and total service time (the combined duration of online and offline service times). Their findings revealed a negative correlation between online service time and workload, with a one standard deviation increase in workload leading to an average decrease of 20.5 seconds in online service time. Notably, the speedup effect during the online interaction was more evident among agents who possess high intrinsic motivation but low extrinsic motivation traits. The behavior of agents during other call stages, which exhibited different patterns, will be discussed in subsequent sections of this paper.

Together, these studies provide evidence that servers, including healthcare providers, bank workers, and call center agents, tend to reduce service time as work-

load increases. Mechanisms such as *rushing*, *task reduction*, *aggressive discharge strategies*, and *motivation* may explain how the pressure of high workloads leads to faster service rates across diverse service environments. This evidence motivates the following hypothesis:

H1: *As workload increases, service time decreases.*

Workload: Slowdown

Contrary to the speedup hypothesis, substantial empirical evidence also supports a slowdown effect, where increased workload leads to longer service times. For example, in a study of 26 emergency department (ED) physicians who handled 145,935 ED patient encounters, Kc (2013) found that higher levels of workload were associated with longer throughput times and lower throughput rates. Similarly, Batt and Terwiesch (2016) have examined the treatment of ED patients suffering from abdominal pains. The authors tested the impact of three measures of workload on patient length of stay (LOS) and service time, namely: ED census (total number of patients in the ED), treatment census (number of patients in the treatment phase), and waiting room census (number of patients in the waiting room). They found a positive correlation between ED/treatment census and LOS/treatment time. However, the relationship between waiting room census and service time was non-linear as discussed in the next section. Similar patterns were observed in a study by Ashkanani et al. (2022), where call center agents demonstrated a positive relationship between total service time and workload. Specifically, a one standard deviation increase in workload corresponded to an 8.4-second increase in total service time. The slowdown effect was more pronounced among agents characterized by low intrinsic motivation and high extrinsic motivation traits. Finally, as discussed earlier, Kc et al. (2020) suggested that ED system load is associated with lower short-term physician productivity.

In summary, while a substantial body of evidence supports the speedup hypothesis as previously discussed, the findings from various sectors including healthcare and call centers also reveal an alternative pattern. Research by Batt and Terwiesch (2016) and Kc (2013), alongside the study by Ashkanani et al. (2022), suggested that increased workload may lead to prolonged service times. These results indicate that the relationship between workload and service time is complex and can vary significantly depending on factors such as the setting and individual level factors of service providers. This variability leads us to consider

a competing hypothesis, reflecting the possibility that increased workload might actually extend service durations rather than reduce them.

H2: *As workload increases, service time increases.*

Workload: Non-linear Effect

Although the evidence presented supports both increases and decreases in service times with rising workload, further investigation suggests a more complex relationship. In particular, the relationship between workload and service time may exhibit non-linear patterns, with the dynamics of service time changing as workload passes certain thresholds. For instance, Tan and Netessine (2014) have examined the reaction of human servers to workload in a restaurant setting. Examining over 190,000 check-level restaurant transactions, the authors found that when workload levels are small, restaurant servers expend effort to increase service quality (as measured by sales) but at the expense of longer service time (as measured by meal duration) since customer waiting cost is low. However, when workload levels become too high, servers try to reduce service time at the expense of service quality to reduce the higher waiting costs. These findings indicate that the relationship between workload and service time is curvilinear, forming an inverted U-shape.

Similarly, in an ED setting, Batt and Terwiesch (2016) found that service time had an inverted U-shape response to workload, specifically to waiting room census. The authors suggest multiple mechanisms that may influence the speedup and slowdown reactions to workload, namely *early task initiation* (i.e., nurses ordering labs at triage as workload increases) and *rushing* (i.e., reduction in nurse-patient interactions as workload increases), both of which are associated with a speedup effect, and *multitasking*, which is associated with an increase in service time. In a similar vein, in a call center setting, Ashkanani et al. (2022) found that the relationship between workload and offline service is curvilinear with an inverted U-shape. The inflection point was approximately at 2 standard deviations above the mean workload levels.

Berry Jaeker and Tucker (2017) suggested that the relationship is even more complex in an inpatient hospital setting. Analyzing over 240,000 patient visits, the authors propose that the relationship between workload and service time is N-shaped. They argued that the effect of workload on workload requirements accounts for this N-shaped pattern. Specifically, at higher levels of workload,

patients with less severe conditions are discharged earlier to alleviate congestion, resulting in a higher proportion of remaining patients with more severe conditions, who in turn need longer service times. Finally, in a Chinese supermarket setting, Wang and Zhou (2018) tested both linear and non-linear workload effects on service time (as measured by cashier checkout time) and found support for a negative linear and convex decreasing relationship between instantaneous workload and service time.

The varied evidence highlights the importance of a flexible approach to understanding workload dynamics. Thus, to capture the complexity of these relationships, two competing hypotheses are proposed: one suggests that servers initially speed up as workload increases until a certain point, after which service times begin to increase; the other suggests that servers initially slow down, with service times decreasing only after workload surpasses a specific threshold. These hypotheses aim to reflect the different initial reactions and adjustments to rising workloads.

H3(a): *Servers initially speed up as workload increases, reducing service times up to a threshold, beyond which workload causes a slowdown.*

H3(b): *Servers initially slow down as workload increases, increasing service times up to a threshold, beyond which workload causes a speedup.*

Overwork: Slowdown

In contrast to workload, the effect of overwork on service time has received relatively scant attention in the behavioral queueing literature. Overwork, conceptualized as a prolonged and excessive workload (Delasay et al., 2019), is frequently linked to an increase in service times, a phenomenon primarily attributed to server fatigue (Kc & Terwiesch, 2009). For instance, Kc and Terwiesch (2009) observed that in healthcare settings, a one standard deviation increase in overwork was correlated with a 1.9% rise in patient transport time and a 17.3% increase in patient length of stay. Similarly, the results found by Staats and Gino (2012) suggested that bank workers exhibited a 5% increase in service time for each standard deviation increase in overwork. These trends are consistent with the findings of Ashkanani et al. (2022), who examined the effect of overwork on call center agents' service times. The analysis revealed that a one standard deviation increase in overwork led to a 10.12-second rise in "online" service time. Interestingly, the evidence for the impact of overwork on offline and total service times was less

clear; the baseline models supported the slowdown effect, while the instrumental variables models suggested speedup and flat effects on "offline" and "total" service times, respectively.

Despite some variations in specific effects, the prevailing evidence from these studies generally supports a slowdown effect of overwork on service time across various industries. This observation forms the basis for the following hypothesis:

H4: *As overwork increases, service time increases.*

Research Setting and Data

The research setting for this study is a call center situated in Kuwait, which operates as part of a large Kuwaiti private hospital. This call center utilizes a pooled queue structure and a next available agent routing rule, ensuring efficient distribution of incoming calls to the first accessible agent. The center provides round-the-clock services, addressing a wide range of patient requests, including appointment scheduling, prescription refills, medical test results, billing and insurance inquiries, and doctor referrals. To maintain continuous operation, the center employs a team of 23 agents working on a shift system. Shift assignments are designed to cover the entire day, with agents rotating through different schedules and days as required. Figure 1 presents the shift assignments for a representative day at the call center, highlighting the distribution of agents during peak and off-peak hours.

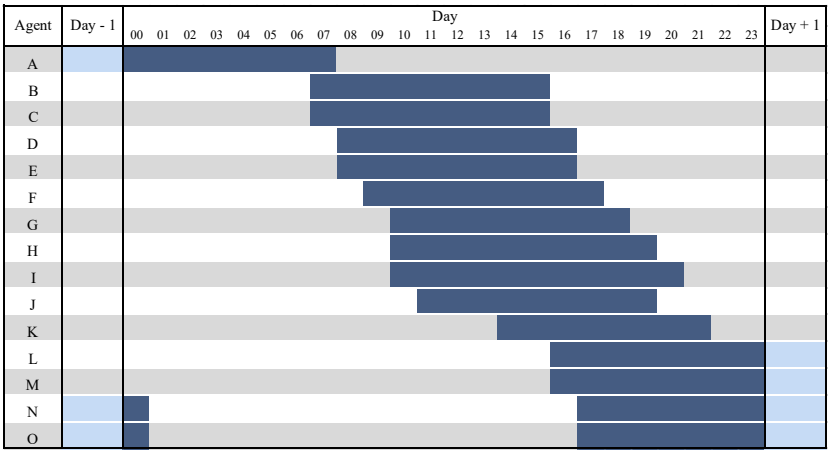

Figure 1: Shift Assignments for a Representative Day

Notes: Dark blue regions denote agent availability within specified time frames for a given day, while light blue areas signify agent working hours that either carry over from the previous day or extend into the subsequent day.

Hypotheses 1- 4 are tested using archival data gathered from the information system of the call center. The dataset contains information on all customer calls handled by a team of 23 agents in October 2016. This represents a total of 59,691 calls processed during that month. Call details include the date and time of the call, the identifier of the agent who processed it, and wait and service time components. Calls with unusually short or long talk times (i.e., top and bottom percentiles) were excluded, resulting in 58,460 calls and 23 agents being included in the analysis. The study uses the notation of "call *i*" and "agent *j*" to refer to specific calls and agents, respectively. Additionally, the data is divided into 30-minute time intervals, referring to the time period in which a call was received. For example, a call received at 8:27 a.m. would be in the 8:00 a.m. - 8:30 a.m. time period.

Measures

Table 2 presents a summary of the variables used in this study. The operationalization of these measures, which were generated using the full sample, is further explained below. Additionally, a data dictionary of the main variables is displayed in Table 3.

Table 2
Descriptive Statistics and Correlations for Main Variables

Variable	Mean	SD	*N*	1	2	3	4
1. Talk Time	118.53	78.72	58.460	-			
2. Call Volume	63.74	23.44	58.460	0.07	-		
3. Workload	9.19	1.95	58.460	0.01	0.56	-	
4. Overwork	-0.69	2.61	58.460	0.02	0.34	0.35	-
5. No. of Agents	6.86	2.23	58.460	0.08	0.88	0.16	0.25

Notes: All correlations with an absolute value > 0.01 are significant at the 1% level.

Dependent Variable

Talk Time ($Talk_{ij}$). I operationalize service time, a common inverse proxy for productivity (Delasay et al., 2019), using talk time, which is measured as the number of seconds an agent spends processing a customer's request while they are on the call. This information is retrieved from the call center's database, which records the time the call starts and ends. The difference between these two values is used to calculate the talk time. It is important to note that the queue waiting time is not included in the calculation of talk time.

Independent Variables

Workload (WL_i). I define *workload* as the ratio of the total number of calls received during a specific time period ($Volume_i$) to the number of agents assigned to work during that time period (NA_i). *Workload* is represented mathematically as $WL_i = \frac{Volume_i}{NA_i}$. This measure of load accounts for the number of agents servicing customers during a specific time period and reflects the average load per agent at the time period that the call was received in (Ashkanani et al., 2022). For example, if the call center received 50 calls during a time period with 5 scheduled working agents, then the workload for that time period would be 10 calls per agent. This way of calculating workload allows us to have a more accurate representation of the real-world situation where the number of agents working during a time period also affects the amount of work that each agent has to handle. The unit of measurement for WL_i is "calls/agent.half-hour".

Overwork (OW_{ij}). I operationalize the *overwork* measure to capture the extended load seen by agent *j* over the two-hour period prior to the arrival of call *i*. This measure of overwork is computed only if agent *j* handling call *i* was on shift during each of the four 30-minute periods prior to receiving the call. Mathematically, *overwork* is defined as $OW_{ij} = \frac{1}{4} \sum_{k=t(i)-4}^{t(i)-1} (WL_k - \overline{WL})$, where *t(i)* denotes the time period call *i* was received in, WL_k denotes workload during time period *k*, and $\overline{WL}$ denotes the grand mean of workload. For example, if an agent receives a call at 10:34 a.m. and the observed workloads during the 30-minute periods starting at 8:30 a.m, 9:00 a.m., 9:30 a.m., and 10:00 a.m. were 5, 6, 7, and 8 calls/agent. half-hour, respectively, and the average workload level was 9.19 calls/agent.half-hour, then the overwork at the time of the focal call's arrival is $\frac{(5 + 6 + 7 + 8)}{4}$ - 9.19 = -2.69, which is 2.96 units below the average workload levels.

Controls

The control variables include both the day-of-week and the time period fixed effects (F.E.). The fixed effects are controlled to account for the variability in the dependent and independent variables that can arise due to differences in operational dynamics across different days of the week and various times of the day.

Table 3
Data Dictionary

Variable	Unit	Description
Talk Time (*TALK*)	Seconds	Number of seconds a call center agent spends processing a customer's request during a call, excluding queue waiting time (used as an inverse proxy for productivity).
Call Volume (*VOLUME*)	Calls/½-hour	Total number of calls received in a 30-minute time period when a focal call is received.
Workload (*WL*)	Calls/ agent.½-hour	Average workload per agent during a specific time period, calculated by dividing the total number of calls received by the number of agents on duty.
Overwork (*OW*)	Calls/ agent.½-hour	A measure of extended load capturing agent's load over the two hours prior to the arrival of a focal call.
No. of Agents (*NA*)	Agents	Number of agents working in the call center during the time period in which a focal call arrived.
Agent Fixed Effects (D)	–	Binary variables capturing individual agent attributes, with 1 denoting the presence of a specific agent's effect and 0 otherwise.
Day-of-week Effects (DOW)	–	Binary variables representing each day of the week, with 1 indicating the presence of a specific day's effect and 0 otherwise.
Time Period Effects (PERIOD)	–	Binary variables representing distinct time periods, with 1 signifying the presence of a specific period's effect and 0 otherwise.

Econometric Specification

Identification Strategy

One potential source of endogeneity in this study is the exclusion of unobserved factors that affect workload, overwork, and service time, which may lead to biased estimates. For instance, an agent's inherent ability or motivation could affect their exposure to load and productivity. In other words, a call center man-

ager might assign more productive agents to days or time periods with heavier call traffic, impacting both the independent variables (i.e., workload and overwork) and the dependent variable (i.e., service time). This concern is mitigated by utilizing a combination of panel data and fixed effects models. The study uses longitudinal data with observations available for agents across several days and time periods, allowing fixed effects models to control for unobserved time-invariant factors that could affect both explanatory variables and the outcome variable.

For example, incorporating agent-specific fixed effects helps control for unobserved time-invariant individual characteristics correlated with both load and service time, such as innate ability, motivation, or personality traits. This helps in addressing omitted variable bias, a common source of endogeneity (Lu et al., 2018). Fixed effects models essentially compare an agent's service time when they experience different levels of workload or overwork, controlling for these time-invariant factors (Allison, 2009; Wooldridge, 2010).

Furthermore, day-of-week effects are included in the models to control for unobserved factors that systematically vary by weekdays, potentially correlating with both load and service time. For instance, workload or overwork might be higher on certain weekdays, as demonstrated in Figure 2. The day-of-week effect could influence both load levels and agent service time. For example, Fridays in Kuwait hold both religious and social significance, with people expected to attend the Jumu'ah prayer and listen to the sermon at the mosque. Additionally, people often spend time with their families, engage in social activities, and attend gatherings with friends and relatives on Fridays. Consequently, call center agents might be less focused and motivated to work on Fridays, leading to decreased productivity. Call volume on Fridays is also lower for the same reasons (i.e., customers are less likely to call since they are busy with their religious and social obligations). Thus, Fridays may affect both agent service time (the outcome variable) and the level of work stressors (i.e., workload and overwork).

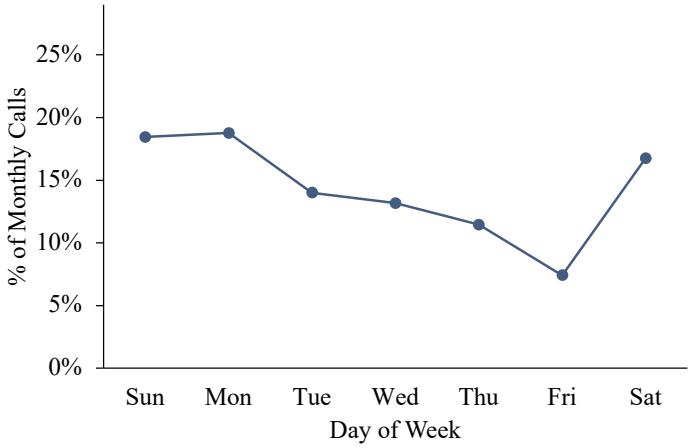

Figure 2: Distribution of Call Volume by Day of the Week

Lastly, time period effects are incorporated into the models to account for unobserved time-of-day effects, which might be systematically correlated with both service time and load. For instance, agents may feel more productive during the day rather than in the afternoon or at night. However, call volume is also correlated with the time of day, as demonstrated in Figure 3. Therefore, controlling for time period effects helps mitigate these concerns. Moreover, the inclusion of various work stress measures can also address some of these issues. Although certain time periods are busier than others, agents are scheduled to work in the call center at different times, as illustrated in Figure 1. Consequently, while agents working during a specific day and time period experience the same workload levels, their overwork measure is expected to vary, as this measure depends on the time they started their shifts.

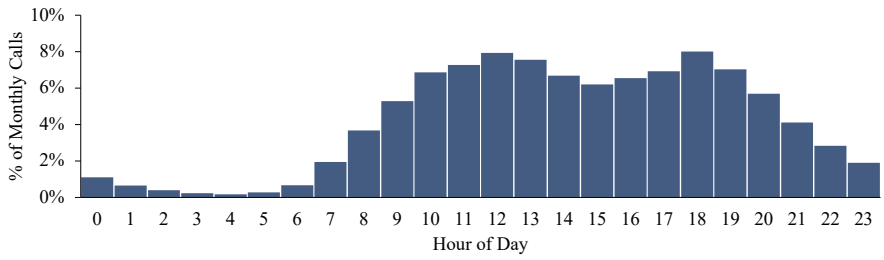

Figure 3: Distribution of Call Volume by Hour of Day

Empirical Models

I test my hypotheses by estimating a series of Ordinary Least Squares (OLS) regression models. The econometric specifications I used to test the hypotheses are listed below. Let X_i denote a vector of control variables, D_j denote a set of dummy variables where $D_j = 1$ for agent *j* (zero otherwise), WL_i and OW_{ij} denote the work stressors of interest, $\overline{WL}$ and $\overline{OW}$ indicate the grand mean of the specified work stressors, and ϵ_{ij} indicate the disturbance term.

a. Aggregate Effects of Workload and Overwork on Service Time

First, I test the aggregate average effect of workload and overwork on service time as a baseline. Models (1) and (2) are used for this purpose. Model (1) includes only the linear effects of workload and overwork, while model (2) includes the quadratic effect of workload to test for non-linear relationships.

$$Talk_{ij}=\alpha_0+\alpha_1 (WL_i - \overline{WL})+\alpha_2 (OW_{ij} - \overline{OW})+D_j A_3+X_i A_4+\epsilon_{ij} \quad (1)$$

$$Talk_{ij}=\beta_0+\beta_1 (WL_i - \overline{WL})+\beta_2 (WL_i - \overline{WL})^2+\beta_3 (OW_{ij} - \overline{OW})+D_j B_4+X_i B_5+\epsilon_{ij} \quad (2)$$

b. Agent-Specific Effects of Workload and Overwork on Service Time

After establishing the baseline effects, I test the hypotheses at an individual level to examine the extent to which each agent's behavior aligns with the hypotheses. This involves including interaction terms between agent fixed effects and the independent variables. The models are specified as follows:

$$Talk_{ij}= \gamma_0+\gamma_1 (WL_i - \overline{WL})+\gamma_2 (OW_{ij} - \overline{OW})+D_j \Gamma_3+(WL_i - \overline{WL})\ D_j \Gamma_4+ (OW_{ij} - \overline{OW})\ D_j \Gamma_5+X_i \Gamma_6+\epsilon_{ij} \quad (3)$$

$$Talk_{ij}= \mu_0+\mu_1 (WL_i - \overline{WL})+\mu_2 (WL_i - \overline{WL})^2+\mu_3 (OW_{ij} - \overline{OW})+D_j M_4+ (WL_i - \overline{WL})\ D_j M_5+(WL_i - \overline{WL})^2\ D_j M_6+(OW_{ij} - \overline{OW})\ D_j M_7+X_i M_8+\epsilon_{ij} \quad (4)$$

The models are estimated using the Stata 18 SE *reg* procedure with clustered standard errors.

Results

Aggregate Impact of Workload and Overwork on Service Time: Baseline Model Results

The results of the baseline regression analyses for models (1) and (2) are presented in Table 4. First, testing the competing Hypotheses 1 and 2, the results of

model (1) demonstrate support for Hypothesis 1, as evidenced by the negative coefficient of workload ($\alpha_1 = -0.587$, $p < 0.001$). Conversely, Hypothesis 4 does not receive support; the effect of overwork is statistically non-significant ($\alpha_2 = -0.189$, *ns*). To evaluate Hypotheses 3a and 3b, I examine the results of model (2), which includes both linear and quadratic terms for the workload effect. The results reveal a statistically non-significant quadratic workload term ($\beta_2 = -0.087$, *ns*) and reaffirm support for Hypothesis 1 with the negative coefficient of the linear workload term ($\beta_1 = -0.589$, $p < 0.001$). Additionally, the statistically non-significant slope of overwork ($\beta_3 = -0.193$, *ns*) indicates a lack of support for Hypothesis 4.

These findings establish a baseline understanding of the grand average effects of workload and overwork on talk time. However, as discussed earlier, these grand average effects might be biased by the composition of human workers who constitute the sample of servers working in the service center. Thus, the next phase of the analysis will shift its focus to the individual agent level, aiming to examine the extent to which individual agent behaviors align with the stipulated hypotheses.

Table 4
Average Effects of Workload and Overwork on Talk Time (Baseline)

	Model (1)	**Model (2)**
Intercept	123.217***	123.537***
	(1.422)	(1.497)
Workload	-0.587***	-0.589***
	(0.159)	(0.152)
Workload^2		-0.087
		(0.057)
Overwork	-0.189	-0.193
	(0.207)	(0.204)
Agent F.E.	Yes	Yes
Time Period F.E.	Yes	Yes
Day-of-Week F.E.	Yes	Yes
No. of Calls	58,460	58,460

Notes: *** denote statistical significance at the 1% confidence level. F.E. denote fixed effects. Clustered standard errors are shown in parentheses.

Agent-Specific Effects of Workload and Overwork on Service Time: An Individual-Level Analysis

Hypotheses 1-4 are tested at the individual level using models (3) and (4). These models include interaction terms that allow us to examine the reaction of each individual server to workload and overwork. In examining the impact of workload on service time across individual agents, the results of model (3) revealed a spectrum of responses that varied significantly. First, a subset of agents exhibited a negative reaction to increased workload, aligning with Hypothesis 1. For these agents, service time decreased as workload intensified. The magnitude of this effect varied. Conversely, consistent with Hypothesis 2, another group of agents showed a positive reaction where service time increased with workload. Interestingly, there was also a notable segment of agents for whom the reaction to workload was essentially flat. This group did not exhibit significant changes in service time as workload varied. These results are illustrated in Figure 4 showing the marginal effects of workload on talk time for individual call center agents.

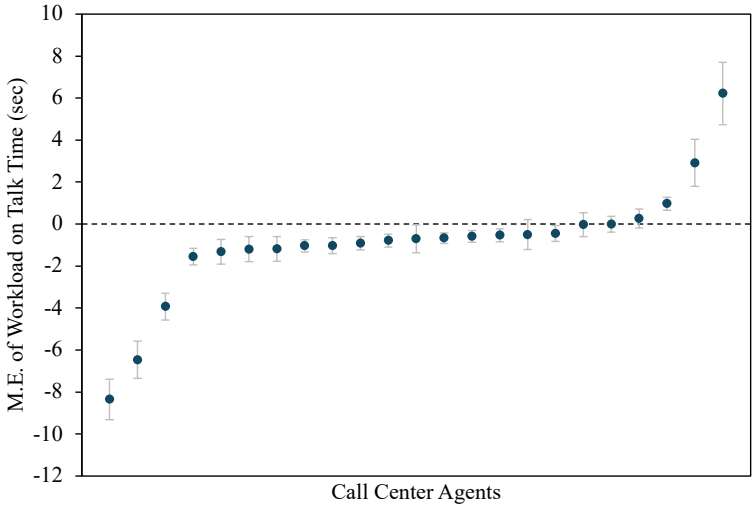

Figure 4: Impact of Workload on Call Center Agents' Talk Time

Notes: Estimates (with 95% confidence limits) of the marginal effect of workload on the talk time of individual call center agents (displayed in ascending order).

In examining the impact of overwork, some agents exhibited behavior consistent with Hypothesis 4, slowing down in response to increased overwork. Howev-

er, other agents showed the opposite effect or had a flat response to overwork, as illustrated in Figure 5. Thus, not all agents showed a behavior consistent with the predictions of Hypothesis 4.

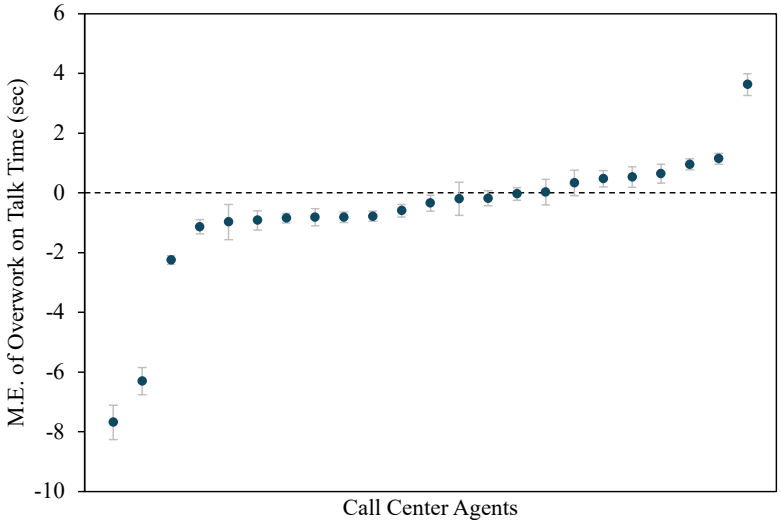

Figure 5: Impact of Overwork on Call Center Agents' Talk Time

Notes: Estimates (with 95% confidence limits) of the marginal effect of overwork on the talk time of individual call center agents (displayed in ascending order).

The results of model (4) allow us to examine non-linear reactions to workload. The findings reveal diverse behaviors among agents, each representing different hypotheses. To illustrate this, I have plotted estimates of the average talk times of four select agents in response to varying levels of workload (see Figure 6). For example, the behavior of the agent in panel (a) is consistent with the predictions of Hypothesis 1, where the agent speeds up as workload increases, dropping the agent's average talk time by 8 seconds per call as workload increases from low to high levels. In contrast, the behavior of the agent depicted in panel (b) is consistent with the predictions of Hypothesis 2. This agent slows down as workload increases, increasing the agent's average talk time by over 11 seconds per call as workload increases from low to high levels. The results in panel (c) are consistent with the predictions of Hypothesis 3a, where the agent initially speeds up with the increase in workload but slows down as workload exceeds average levels. Finally, the results in panel (d) show the opposite behavior, which is consistent with the predictions of Hypothesis 3b. This agent initially slows down with the increase in

workload when the overall workload levels are low but starts speeding up when workload levels exceed the average workload. In summary, the results of model (4) suggest the following distribution of agent behaviors: 9% were consistent with Hypothesis 1, 4% with Hypothesis 2, 26% with Hypothesis 3a, and 61% with Hypothesis 3b.

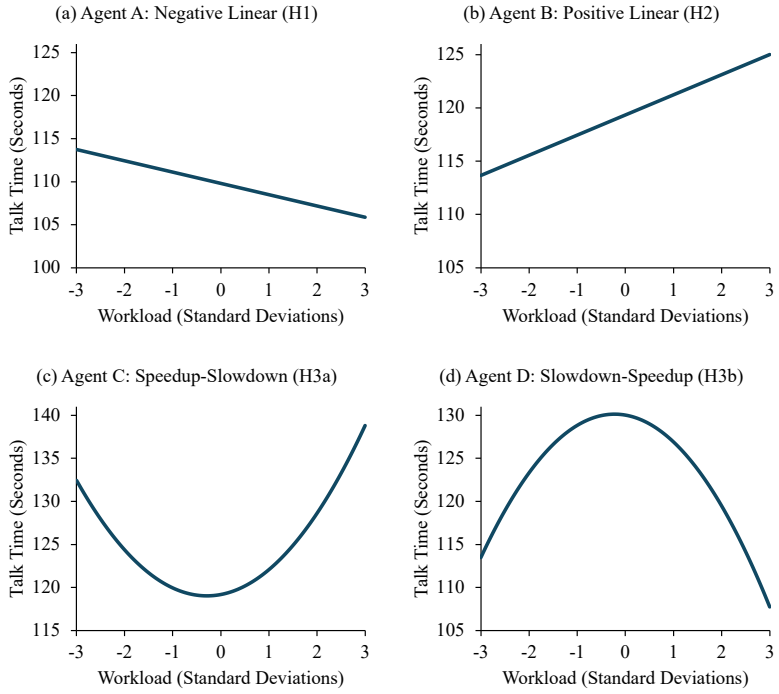

Figure 6: Agent Effect on Reactions to Workload

Notes: Estimates of average talk time (y-axis) for select agents as a function of workload (in standard deviations, x-axis). Each panel represents the behavior of a single agent, illustrating different observed reactions to workload within the sample.

Robustness Check: Overwork Variants

Overwork is a measure of extended load (Delasay et al., 2019), which captures an agent's workload experience during a specified time frame prior to receiving a focal call. In other words, if call *i* was received after a sustained period of high (low) agent load, then the overwork measure experienced by agent *j* will be positive (negative). Following Ashkanani et al. (2022), overwork is operationalized by calculating the average workload over a 2-hour interval prior to receiving call *i*.

To further validate the results, I assessed the impact of using different time periods for overwork measurement. Specifically, I estimated model (3) with various overwork windows, including 1-hour, 1.5-hour, 2.5-hour, and 3-hour time frames. The results from estimating these alternative versions of model (3) are presented in Figure 7, with subplots (a) through (d) illustrating the estimated coefficients for the marginal effect of overwork on talk time across different overwork variants. The findings are consistent with the original analysis, showing that agents respond differently to overwork–some speed up, others slow down, and some show no reaction. This confirms the original observation of varying sensitivities and reactions to overwork among agents.

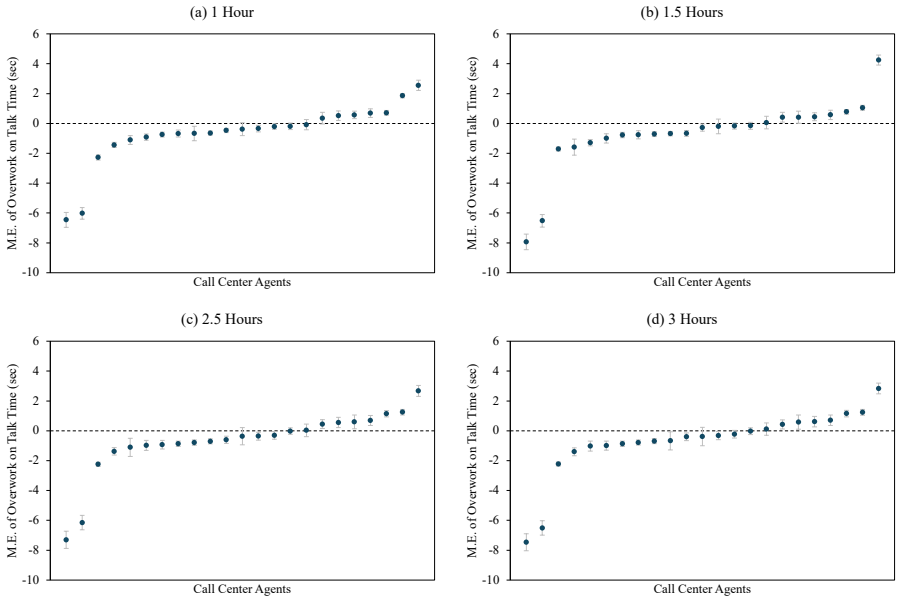

Figure 7: Impact of Overwork Variants on Call Center Agents' Talk Time

Note: Estimates (with 95% confidence limits) of agent reaction to overwork variants are displayed for the call center agents in ascending order (order is based on the estimates of each subplot), with overwork defined using (a) 1-hour, (b) 1.5-hour, (c) 2.5-hour, and (d) 3-hour windows

Discussion

This study investigated the variations in reactions to workload and overwork among call center agents. The results revealed a spectrum of responses: while

some agents exhibited a speedup effect, reducing service time under increased workload, others showed a slowdown effect, where service time increased. Additionally, a significant portion of agents demonstrated non-linear responses, either speeding up then slowing down or slowing down then speeding up beyond a certain workload threshold. The impact of overwork similarly varied, with some agents slowing down in response to higher overwork levels, while others had flat or even speedup responses. These findings underscore the complexity of human behavior under different work stressors, highlighting the need for tailored strategies to optimize productivity in call centers.

Theoretical Contributions

The findings of this study support the behavioral queuing literature's assertion that server productivity behavior is complex, endogenous to work stressors, and subject to a variety of contingency factors. Previous studies have identified a range of workload impacts on service time, including speedup effects (Ashkanani et al., 2022; Freeman et al., 2017; Kc & Terwiesch, 2009, 2012; Kc et al., 2020; Staats & Gino, 2012), slowdown effects (Ashkanani et al., 2022; Batt & Terwiesch, 2016; Kc, 2013), and non-linear effects (Ashkanani et al., 2022; Batt & Terwiesch, 2016; Berry Jaeker & Tucker, 2017; Tan & Netessine, 2014; Wang & Zhou, 2018). This study contributes to the behavioral queuing literature by developing and testing an individual-level model that captures agent-specific responses to work stressors.

The empirical results of this study demonstrate the advantage of this individual-level approach, revealing insights that are often obscured by conventional aggregate-level analyses of human servers' reactions to work stressors. For instance, if a sample is dominated by servers who speed up in response to workload, the overall average response will likely indicate a speedup effect, and vice versa. This was empirically observed in our study, where the aggregate workload effect was negative, indicating a linear speedup behavior in response to higher workload. However, an individual-level analysis revealed diverse behaviors, including various linear and non-linear reactions, which were masked by the aggregate analysis.

Furthermore, if a sample consists of a balanced mix of opposite behaviors (e.g., 50% speedup and 50% slowdown), the overall effect may appear as a flat response to a work stressor. This aggregate effect can be misleading, as it does not reflect the fact that agents might be sensitive to changes in work stressor levels but have opposite directional responses. This was observed in this study with

the aggregate overwork effect, which appeared flat (statistically non-significant). However, individual-level analysis revealed that most agents responded to overwork, but in opposite directions.

By developing and testing individual-level models, this study provides a nuanced understanding of the true nature of responses to workload and overwork. It highlights the importance of examining individual-level behaviors, which can yield deeper insights and inform more effective management strategies in service operations. This approach aims to advance the behavioral queuing literature by offering a detailed and accurate depiction of server behavior under varying work stressors. Moreover, the study offers a potential explanation for the mixed results observed in the behavioral queuing literature regarding the workload-service time relationship. It suggests that these inconsistencies might be influenced, at least in part, by differences in the composition of human servers across various studies. This is particularly relevant considering that empirical queuing studies often involve a relatively small number of servers (typically 20-100 per study) compared to the large number of transactions these servers process (ranging from thousands to hundreds of thousands per study). By accounting for variations in servers' responses to work stressors, this research aims to provide a clearer and more comprehensive understanding of the dynamics at play, thereby contributing to a more refined body of knowledge in the field.

Practical Implications

The findings from this research offer valuable guidance for call center managers aiming to enhance their center's performance. These insights are particularly useful for optimizing agent shift allocation. With a finite number of agents available, it is crucial for managers to aim for a balanced mix of agents who have different responses to work stressors within each shift. For example, a shift consisting only of agents who tend to slow down when stressed could lead to reduced call center productivity and diminished customer service quality. Conversely, a blend of agents who slow down and those who speed up under stress can create a more balanced and effective work environment.

Additionally, call center managers can provide targeted training and support tailored to agents' productivity patterns and stress responses. Agents who accelerate their work pace under stress might benefit from stress management training to prevent burnout, along with quality management training to avoid compromises

in service quality. On the other hand, agents who tend to slow down might need motivational coaching to boost productivity during peak times.

In summary, by understanding and applying the insights from agent-specific reactions to work stressors, call center managers can significantly improve overall performance. This approach leads to more strategic shift allocation, focused training and support, and ultimately enhances the overall performance of the call center.

Limitations and Directions for Future Research

Despite its contributions, this study has several limitations that open avenues for future research. First, while the number of observations was large, with tens of thousands of transactions, these transactions were handled by a relatively small sample of only 23 call center agents. This may limit our understanding of the prevalence of diverse reactions to workload and overwork among service agents. Future research should aim to solicit data from a larger number of providers to enhance the generalizability of the findings. However, it is worth noting that while the sample of human servers was relatively small, it revealed a full spectrum of predicted behaviors, ranging from speedup to slowdown, and even non-linear or flat responses to work stressors. This suggests that diverse reactions are indeed present even in this relatively smaller sample. Additionally, other studies in the literature often feature a relatively small number of servers or do not report the number of servers at all. Therefore, as a general practice, scholars should report the number of individual servers, not just the number of observations in a transactional dataset.

Another limitation is that while this study showed the diversity in responses to work stressors among call center agents, it did not explore the individual-level mechanisms driving these differences. For example, Ashkanani et al. (2022) suggest that intrinsic and extrinsic motivations are linked to differences in responses to workload among call center agents. Other factors such as personality, ability, or additional motivational mechanisms might also play a role, and future research should explore these. Nonetheless, the individual-level approach proposed in this study is particularly useful when measuring individual differences in servers' traits or states is not feasible. In such cases, characterizing the distribution of responses rather than just examining the aggregate effect can provide valuable insights.

Future studies could also attempt to replicate this study across different contexts. This includes call centers with varied queue structures (e.g., shared vs. ded-

icated queues), different compensation schemes (e.g., fixed salaries vs. performance-based pay), and in various service industries such as healthcare, banking, or restaurant services. Exploring these different settings would help validate the findings and potentially uncover additional factors influencing server responses to workload and overwork, thereby broadening the applicability and robustness of the theoretical contributions presented here.

References

Allison, P. D. (2009). *Fixed effects regression models*, SAGE Publications. https://doi.org/10.4135/9781412993869

Allon, G. & Kremer, M. (2018). *Behavioral foundations of queueing systems*. The Handbook of Behavioral Operations, 323-366. https://doi.org/10.1002/9781119138341.ch9

Ashkanani, A. M., Dunford, B. B., & Mumford, K. J. (2022). Impact of motivation and workload on service time components: An empirical analysis of call center operations. *Management Science, 68*(9), 6697–6715. https://doi.org/10.1287/mnsc.2022.4491

Batt, R. J., & Terwiesch, C. (2016). Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science, 63*(11), 3531–3551. https://doi.org/10.1287/mnsc.2016.2516

Bendoly, E., & Prietula, M. (2008). In "the zone": The role of evolving skill and transitional workload on motivation and realized performance in operational tasks. *International Journal of Operations & Production Management, 28*(12), 1130–1152. http://dx.doi.org/10.1108/01443570810919341

Berry Jaeker, J. A., & Tucker, A. L. (2017). Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science, 63*(4), 1042–1062. https://doi.org/10.1287/mnsc.2015.2387

Delasay, M., Ingolfsson, A., Kolfal, B., & Schultz, K. (2019). Load effect on service times. *European Journal of Operational Research, 279*(3), 673–686. https://doi.org/10.1016/j.ejor.2018.12.028

Freeman, M., Savva, N., & Scholtes, S. (2017). Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science, 63*(10), 3147–3167. https://doi.org/10.1287/mnsc.2016.2512

Kc, D. S. (2013). Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management, 16*(2), 168–183. https://doi.org/10.1287/msom.2013.0464

Kc, D. S., Staats, B. R., Kouchaki, M. & Gino, F. (2020). Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science, 66*(10), 4397–4416. https://doi.org/10.1287/mnsc.2019.3419

Kc, D. S., & Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science, 55*(9), 1486–1498. https://doi.org/10.1287/mnsc.1090.1037

Kc, D. S., & Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management, 14*(1), 50–65. https://doi.org/10.1287/msom.1110.0341

Lu, G., Ding, X. D., Peng, D. X., & Chuang, H. H. C. (2018). Addressing endogeneity in operations management research: Recent developments, common problems, and directions for future research. *Journal of Operations Management, 64*(1), 53–64. https://doi.org/10.1016/j.jom.2018.10.001

Shunko, M., Niederhoff, J., & Rosokha, Y. (2017). Humans are not machines: The behavioral impact of queueing design on service time. *Management Science, 64*(1), 453–473. https://doi.org/10.1287/mnsc.2016.2610

Song, H., Tucker, A. L., & Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science, 61*(12), 3032–3053. https://doi.org/10.1287/mnsc.2014.2118

Staats, B. R., & Gino, F. (2012). Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science, 58*(6), 1141–1159. https://doi.org/10.1287/mnsc.1110.1482

Tan, T. F., & 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. https://doi.org/10.1287/mnsc.2014.1950

Wang, J. & Zhou, Y. P. (2018). Impact of queue configuration on service time: Evidence from a supermarket. *Management Science, 64*(7), 3055–3075. https://doi.org/10.1287/mnsc.2017.2781

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Ahmad M. Ashkanani is an Assistant Professor in the Information Systems and Operations Management Department at Kuwait University's College of Business Administration (CBA). He is the head of the Research Development Unit at CBA. He earned his Ph.D. from the Krannert School of Management at Purdue University. He specializes in behavioral operations, healthcare management, service operations, and technology applications in OSCM. His research has been published in Management Science, Health Care Management Review, and other journals. (a.ashkanani@ku.edu.kw)

Acknowledgments and Disclosures

Author would like to express his gratitude to the anonymous reviewers for their constructive feedback, which helped enhance the quality and contribution of this paper. Additionally, he extends his thanks to Abdullah Alhauli, Salman Aljazzaf, and Yousef Abdulsalam for their valuable comments on earlier versions of the manuscript.