When More is Less: Field Evidence on Unintended Consequences of Multitasking

Paulo B. Goes
Eller College of Management, University of Arizona, Tucson, Arizona 85721, pgoes@eller.arizona.edu

Noyan Ilk
College of Business, Florida State University, Tallahassee, Florida 32306, nilk@business.fsu.edu

Mingfeng Lin
Eller College of Management, Univ. of Arizona, Tucson, Arizona 85721, mingfeng@eller.arizona.edu

J. Leon Zhao
College of Business, City University of Hong Kong, Hong Kong, China, jlzhao@cityu.edu.hk

February, 2017

Abstract

Online customer service chats provide new opportunities for firms to interact with their customers and have become increasingly popular in recent years for firms of all sizes. One reason for their popularity is the ability for customer service agents to multitask, i.e., interact with multiple customers at a time, thereby increasing the system "throughput" and agent productivity. Yet, little is known about how multitasking impacts customer satisfaction, the ultimate goal of customer engagements. We address this question using a proprietary dataset from an S&P 500 service firm that documents agent multitasking activities (unobservable to customers) in the form of server logs, customer service chat transcripts and post-service customer surveys. We find that agent multitasking leads to longer in-service delays for customers, and also lower problem resolution rates. Both lead to lower customer satisfaction, although the impact varies for different customers. Our study is among the first to document the link between multitasking and customer satisfaction, and has implications for the design of agent time allocation in contact centers, and more broadly for how firms can best manage customer relations in new service channels enabled by IT.

Key words: multitasking; customer satisfaction; service operations; information systems: IT policy and management

1

1. Introduction

Modern service organizations are under constant pressure to satisfy ever increasing service demands with limited operating budgets, of which human capital expenses constitute almost 70% (Human Capital Management Institute 2010). Not surprisingly, many service organizations adopt multitasking policies, under which employees handle multiple tasks simultaneously and switch from one task to another on a regular basis (Aral et al. 2012). This seems to be a natural and intuitive decision, when there is idle time during the execution of a particular task. Multitasking allows employees to better utilize their work time; and increase the number of "units" that they can process in a given time frame, thus improving the overall productivity. One of the most prominent applications of multitasking in the workplace can be observed in contact centers, especially in online live chats in which agents regularly shift their attention among different customers. Online customer service live chat has become a new channel for a firm to engage new customers and retain existing customers. It is being used not only in technology startups but also by many traditional service firms, ranging from banks, car dealers, and insurance companies to hospitality, legal, and travel agencies (Reinsch et al. 2008). By putting more customers in touch with an agent at a given period of time, the benefit of multitasking in improving productivity seems obvious.

However, productivity of agents, especially in live chats, is much more than the "throughput" of how many customers an agent can handle. Customer satisfaction (Fornell et al. 1996), a critical factor in customer acquisition and retention, is a well-established aspect of any performance metrics for customer interactions that rarely been addressed in the research and practice of multitasking. Firms will not be able to win new customers or retain existing customers if these customers are "processed" through queues, but not satisfied. If customers are unable to resolve their issues satisfactorily via online visits, they may resort to more traditional means of customer contact such as phone calls or in-person visits (Bavafa et al. 2017), rendering online chat irrelevant. Our goal is to fill this gap in the literature. Specifically, we address the following research question: *How does multitasking affect customer satisfaction in online service live chats?*

2

To investigate this problem, we utilize a comprehensive proprietary dataset from the call center operations of an S&P 500 tax preparation services firm. This contact center receives requests from customers using a tax preparation software and provides tax filing and technical assistance through its human service agents, who communicate with customers via a live-chat interface. The environment is very dynamic with jobs arriving and leaving the system in a relatively short period of time. To increase the "throughput" of the online chat channel, the system automatically assigns multiple chat conversations to service agents. In other words, customers' arrival in the queue and their assignment to agents are both exogenous, therefore, the multitasking level of agents is exogenously determined. This provides a uniquely ideal opportunity to study the impact of chat agent multitasking.

We focus on three important consequences of multitasking that have received relatively little attention in the empirical literature, especially for the chat center context. The first one is *problem resolution rates*. An inherent side effect of multitasking is its burden on cognitive capacity of agents (Charron and Koechlin 2010). Specifically, task switching and interruptions due to multitasking may lead to difficulty with task re-orientation and lower performance (Adler and Benbunan-Fich 2012). We therefore hypothesize that multitasking will lead to lower problem resolution rates. The second consequence we study is *in-service delays* (i.e., response delays). When an agent chats with multiple customers, each customer will not receive as much attention as when they are the sole customer, therefore it will take longer for customers to hear back from agents, resulting in longer in-service delays. Furthermore, and even more importantly, service interruption and delays and lower problem resolution rates will both lead to lower customer satisfaction (Diaz and Ruiz 2002; Tom and Lucey 1995). Given the critical importance of customer satisfaction for organizations (e.g., Fornell et al. 1996), the third and final outcome variable that we study in this paper is *customer satisfaction.*

In our empirical tests, we first examine problem resolution and response delay as a function of multitasking. We measure multitasking in two different ways, one focusing on the fact that an agent is *assigned to* multiple customers, and the other focusing on the fact that an agent is *actively working with*

multiple customers. These two complementary metrics allow us to open the "blackbox" of multitasking and understand how agents behave. We find that agents are less likely to resolve customer problems if they are actively engaged in two or more tasks concurrently. On the other hand, assignment of a new customer alone is sufficient for the agents' responsiveness to decrease, even if the agent is focusing on a single customer. We then turn to the third outcome variable of interest and show that problem resolution and response delay further lead to negative impact of multitasking on customer satisfaction using data from the company's follow-up customer surveys. These results are consistent after controlling for the mechanical effects of multitasking (such as delays solely caused by typing a longer text message). Our results are also robust to alternative specifications such as heterogeneity among agents, time-varying effects and self-selection bias.

We further investigate if such effects are uniform across varying agent workloads and different customer segments. We find that the marginal impact of having an additional customer is increasingly higher. Further, we find that not all customers have the same degree of intolerance towards multitasking. To illustrate this, we use customer-specific variables observable at the time of their arrival and estimate a finite mixture model that accounts for heterogeneous customer responses to the negative impacts of multitasking. This suggests that contact centers can practically improve their customer-routing strategies— which customers should be assigned to multitasking agents, and which should not—to improve the system throughput while minimally reducing customer satisfaction.

Our study contributes to the literature and practice in several important ways. First, our study is the first to address and quantify the relationship between multitasking and customer satisfaction, bridging these two fields that are each important in their own right. Second, we draw on data collected through multiple channels and multiple levels, including session-level and message-level data, as well as post-service customer follow-up survey data that are precisely linked to agents' chat sessions. The extensiveness and comprehensiveness of the data is rare, to the best of our knowledge, in multitasking literature. Third, we go beyond the relationship between multitasking and satisfaction, and address the heterogeneity of such impacts from both the agent's and customer's points of view. Our detailed findings can help managers

develop and improve customer-routing strategies that can reduce the negative impact of multitasking, while retaining its productivity benefits.

2. Related Literature and Hypotheses

There has been a long standing literature on the relationship between workload, multitasking, and productivity (Tan and Netessine 2014). For example, Aral et al. (2012) found that at low levels of multitasking (measured as the average number of projects simultaneously worked on), workers attain benefits from task complementarities and smoothing bursty work, leading to increased output. Cameron and Webster (2011) observed that busy workers may make themselves more accessible to their colleagues via multitasking. Reinsch et al. (2008) argued that multitasking team members in a meeting will communicate more efficiently by initiating concurrent live-chat conversations among themselves. To our knowledge however, there is no existing literature that directly links multitasking to customer satisfaction. To motivate our empirical analyses, we draw on several research streams to develop our hypotheses.

2.1. Multitasking and In-Service Delays

There has been a number of studies that considered the effects of multitasking during the process of a task (i.e. *in-service*). These studies can be broadly categorized into two groups. The first group approaches the problem from a queueing theory perspective and aims to develop analytical models for live-chat contact centers that enable multitasking. For instance, Campello et al. (2016) develop a stochastic model of how agents (case managers) process simultaneous jobs with multiple processing steps. Luo and Zhang (2013) model a contact center as a pool of many homogeneous servers, each operating under the processor-sharing protocol. They provide an asymptotic analysis and a fluid approximation of the system, which may help with staffing and admission control decisions. Tezcan and Zhang (2014) develop an analytical model for a similar system using linear programs (LP). They propose a solution to the staffing LP and prove that this solution is asymptotically optimal. Further, they characterize multitasking as a well-known queueing service discipline known as generalized processor-sharing (Demers, Keshav and Shenkar 1989). Under this discipline, the server's work effort can be shared among the jobs in service using time-sharing mechanisms

(Tan et al. 2005). A common approach to divide the work is to use round robin scheduling, in which time slots are assigned to each job in equal portions and in circular order. The server processes each job in turns and only for the amount of time defined by a job's allocated time slot. Evidently, the more jobs there are that share a single server, the more that these jobs wait idly between their turns[1]. It is important to note that such idle times are simply mechanical consequences of the multitasking activity (due to round robin scheduling) and the studies mentioned within this group do not consider additional task-switching related factors.

On the other hand, the second group of literature emphasizes the switching costs associated with the multitasking activity and aims to specifically account for such costs. For example, Coviello et al. (2014) develop an economic production function that describes the slowdown in the output of a worker due to multitasking. Based on this function, they derive a law that can determine the output rate given the number of workers and the multitasking policy. Hall et al. (2015) define a system in which machines are susceptible to work interruptions and task-switching costs. They include switching costs in their analyses and develop optimal algorithms for single machine scheduling problems with multitasking.

Collectively, these two research streams emphasize the following two characteristics regarding the relationship between multitasking and in-service delays: 1) multitasking causes in-service delays between processing periods due to the nature of round robin scheduling and 2) these delays can be further inflated due to switching costs associated with shifting between different tasks. In our study, we expect to see similar effects in the form of service agent responsiveness. Agent responsiveness in a live-chat session can be defined as the agent's punctuality (i.e., promptness) in replying to individual customer messages. A typical chat conversation consists of multiple text messages being transmitted between the agent and the customer

---

[1] To illustrate, consider a hypothetical example, in which there is a single server with a 100 second total service capacity. If there were two jobs that share this service capacity, each time slot would be 50 seconds and a given job would need to wait for $100 - 50 = 50$ seconds between its processing turns. On the other hand, if there were four jobs that share this service capacity, each time slot would be $100 / 4 = 25$ seconds and a given job would need to wait for $100 - 25 = 75$ seconds between its processing turns.

via a chat interface. These messages are generally short in order for the other participant to respond quickly, thus creating a feeling of face-to-face conversation. According to an industry benchmark survey by Telus International (2011), agent responses in a live-chat session should be provided within 30 seconds to generate a sense of real-timeliness. Otherwise, longer periods of idle time between messages could be considered as in-service delays by customers.

Within the context of live-chat communication, we posit that multitasking will lead to increased in-service delays due to its processor-sharing nature as well as the switching costs associated with it. In particular, we expect the impact on in-service delays to grow non-linearly with the increasing levels of multitasking activity[2]. We therefore hypothesize:

HYPOTHESIS 1 (H1): *Multitasking increases agent response delays in live-chat conversations. The marginal effect of more multitasking on response delays should be increasing.*

2.2. Multitasking and Task Resolution

There exists a rich body of research on the cognitive effects of multitasking including dual-task interference, increased cognitive load, and task accuracy (Cameron and Webster 2013). Psychology literature notes that loss of context associated with switching tasks can be a major delay factor during the resumption of the initial task (Czerwinski et al. 2004). Mark et al. (2008) argue that workers try to compensate for the loss of performance due to interruptions by working faster, which comes at a price of increased stress and frustration. Along with increased multitasking workload, stress and frustration may result in cognitive overload, which leads to processing mistakes. In the end, service quality and worker performance may suffer (KC 2013).

There have been studies that investigated the validity of this argument in experimental settings. Using brain imaging technologies, Charron and Koechlin (2010) show that frontal brain function is vulnerable to

---

[2] This non-linearity is assumed to be a result of additional slowdown due to cognitive load as well as the diminishing marginal benefits of bursty work at increased levels of multitasking.

mistakes when an individual pursues two concurrent goals simultaneously. Adler and Benbunan-Fich (2012) conduct a controlled experiment to compare multitasking and non-multitasking conditions, and find that increased levels of multitasking lead to a significant loss in accuracy. Bailey and Konstan (2006) demonstrate that task interruptions and continuous task switching increase the stress and anxiety levels of individuals and lead to up to twice the number of errors committed across the tasks.

A limitation with these studies is the simplistic and artificial nature of tasks that are simulated within the lab environment. Our paper aims to contribute to this literature by studying multitasking in the context of complex, real-world tasks. Within the contact center environment, we postulate that service agents will be less effective in handling customer problems when they are engaged in multiple conversations at the same time. Accordingly, we hypothesize that:

HYPOTHESIS 2 (H2): *Multitasking reduces the agents' likelihood of resolving customer problems in live-chat conversations.*

2.3. Determinants of Customer Satisfaction

Customer satisfaction is a quintessential cornerstone for service organizations to achieve success and long-term profitability (Anderson et al. 1997). It is a complex construct with a variety of determinants including expectations, disconfirmation, performance, affect and equity (Szymanski and Henard 2001). In the context of contact centers, we argue that in-service delays and task resolution are two of the major determinants of customer satisfaction.

To understand the impact of in-service delays on customer satisfaction, we must first distinguish different types of customer-centric waiting times. A common categorization is to define different types of wait based on the point of time at which the wait is initiated. Customers can wait before the process (*pre-service*), during the process (*in-service*) and after the process (*post-service*) (Dube-Rioux et al. 1989). Taylor (1994) defines *delay* as the pre-service, post-schedule wait, which occurs when a scheduled event doesn't commence on time. Diaz and Ruiz (2002) broaden the concept of delay to in-service wait using a field study of delayed flights at an airport. In both studies, common results indicate that longer delays result

in greater anger among customers, and anger leads to diminishing customer satisfaction with services. Similarly, Chebat and Filiatrault (1993) show that interrupted service (such as commonly seen in multitasking) has a negative effect on the perceived waiting duration, clients' mood and the service quality. Further, all of these studies emphasize the scheduled aspect of delays. A delay is different than waiting in queue in the sense that it is a broken promise; customers are scheduled to receive services but are kept on hold during the process. Response delay in live-chat conversations is a similar form of this situation. Inevitably, customers would expect to receive timely responses to their messages once they have been admitted into service.

A consensus in the literature is that the outcome of a service encounter significantly and directly affects customer perceptions regarding the quality of service (Brady and Cronin 2001). In a seminal study, Zeithaml et al. (1996) find that customers who receive satisfactory resolution of their service problems have significantly higher loyalty and retention intentions, along with increased willingness to pay a premium for the service. They conclude that effective service resolution and recovery significantly improve all facets of customers' behavioral intentions. Later studies have solidified the argument that task resolution incompetency, along with longer in-service delays, would frustrate customers and lead to a decline in overall customer satisfaction. In this study, our goal is to link multitasking to customer satisfaction. We expect that such effects will be indirect—i.e., they will be transmitted via task resolution and in-service delays—since customers cannot directly observe that the agent is multitasking. Accordingly, we hypothesize that:

HYPOTHESIS 3a (H3a): *Multitasking in live-chat conversations indirectly affects customer satisfaction via agent response delays.*

HYPOTHESIS 3b (H3b): *Multitasking in live-chat conversations indirectly affects customer satisfaction via problem resolution.*

In addition to testing these hypotheses, we also develop and estimate a finite mixture model that takes possible customer heterogeneity into account while studying the impact of multitasking. This model has the flexibility to classify customers into segments based on customer-specific characteristics and simultaneously estimates the impact of multitasking on customer satisfaction for individual segments.

3. Empirical Context and Data

3.1. Contact Center

To estimate our models, we utilize multiple sources of data from a large-scale live-chat contact center between 01/03/2011 and 02/02/2011. The contact center is operated by a firm that develops tax preparation and filing support software and services. Many firm customers require customer support for technical problems as well as tax filing assistance. The firm categorizes incoming service requests into 20 distinct skill types and trains each service agent to respond to one or more types. In addition, skill types are classified into 3 broader domains with respect to the context of the conversation. Due to the intricate nature of United States income tax laws and the dire consequences of filing incorrect tax returns, the contact center receives a significant amount of traffic, especially during the tax filing season.

To help meet this demand, service agents multitask whenever there are customers waiting in the queue, with up to 4 parallel chat sessions. The queues are formed at the department level and a particular queue is served by multiple agents. Customer-agent assignments are made automatically[3] by the system on a first-come-first-served basis. Further, multitasking assignments are determined by the system according to the length of the queue. The contact center operates from 5 AM to 10 PM daily. Agents are only active at certain time periods during the day, with 4 hours of work on average each day. Figure 1 depicts the daily and average hourly chat requests in the system.

INSERT FIGURE 1 HERE

---

[3] One exception is transferred cases (i.e., agents can manually transfer a customer to another agent). We exclude all transferred cases from the data set in our analysis since they only account for a very small portion of the data.

3.2. Data

We access three data sources within the contact center and conduct an extensive data preparation and consolidation process in order to construct the measures and variables relevant to our analysis. The first data source is transaction server logs. Also called the meta-data, this set of records includes information about each chat transaction's arrival time to and departure time from the system, queue details, service request specific information such as skill key, information about agent assignments and the amount of time spent in queue as well as in service. This data source is also used as one of the inputs to extract information about the agents' multitasking activities.

The second data source is a repository that stores the actual chat conversations in the system in text form (i.e., transcripts) and initial problem descriptions provided by customers when they arrived in the queue. Chat transcripts consist of all the messages exchanged between the agent and the customer. We analyze these transcripts to extract the time stamp of each message. For each customer message, we compute how long it took for the agent to respond to the customer by taking the difference between time stamps. We also check how many customers the agent was talking to during this response delay to obtain multitasking information[4]. Specifically, we count all open sessions during the delay period as well as the effort of the agent (e.g., number of messages sent and number of words typed). Figure 2 illustrates this idea for a particular message in an illustrative conversation timeline. We also collect customer message lengths and customer response delays prior to agent messages to be used as control variables.

INSERT FIGURE 2 HERE

In addition, chat transcripts are accompanied by a text-based description of the problem and the reason for the contact request. Customers are required to enter this information before they are admitted into the system. Problem descriptions are important because they may reveal the initial mood of customers, which

---

[4] Multitasking information is not stored in the databases of the contact center. Therefore, we develop procedures to measure multitasking levels from the existing data sources.

may influence their eventual satisfaction. We combine all problem description text with the first three messages of the customer and employ sentiment analysis (Pennebaker et al. 2001) on the aggregated text to extract the mood of customers at the beginning of the chat sessions.

The third data source contains the survey response information. Each customer completing a chat session receives an e-mail survey (sent one day after the chat session) from the company with questions about their chat experience, satisfaction and problem resolution. During the period that our data covers, overall survey response rate was 16%. A summary of the information contained in individual data sources is provided in Table 1.

<div align="center">INSERT TABLE 1 HERE</div>

3.3. Data Cleaning

We pre-process the consolidated data to remove the observations that have systematically different chat characteristics than the rest of the sample. First, we find and remove all chat sessions that were transferred between service agents. Second, we find and remove all agent-to-agent chat conversations[5]. While the live-chat system permits agents to communicate with other agents, these conversations have different intent and dynamics than typical customer-agent interactions. Finally, we remove extreme observations from the consolidated data set. A rare but important situation is the system's failure to properly terminate a chat conversation. In such cases, certain variables (e.g. service duration, response delay, and customer slowness) can get artificially inflated. To mitigate this, we exclude observations for time-related variables whose values exceed the 99th percentile. If there exists an extreme observation at the message level, we remove the entire session related to this observation. Another issue is copying and pasting text while crafting responses. In our message-level analysis, we control for the delays associated with the agent's typing

---

[5] Our data cleaning is performed after multitasking computations. We use the entire system data (including agent-to-agent sessions) for multitasking computations because otherwise we may miss certain sessions that were multitasked but preemptively removed from the data set. Once multitasking values are computed, we remove agent-to-agent sessions from the data.

behavior (i.e., mechanical efforts) using the number of words typed. However, on rare occasions, agents copy and paste long chunks of standard text into messages. For such cases, messages lengths do not accurately reflect agents' efforts. To mitigate this, we search for messages that were extremely long (over 99th percentile), yet took less than the 50th percentile of the overall response time. Then, we replace the length of these messages with the mean message length in the dataset. Table 2 shows the contact center summary statistics after data pre-processing and cleaning steps.

INSERT TABLE 2 HERE

3.4. Variable Definitions and Measurement

We conduct our empirical analysis at two different levels: message and session. Message-level analysis is used to investigate the impact of multitasking on individual agent response delays (i.e., in-service delays). Session-level analysis is used for two purposes: 1) to investigate the impact of average session multitasking on problem resolution and 2) to investigate the indirect impact of average session multitasking on customer satisfaction. After data collection and processing, we construct the following variables.

3.4.1. Dependent Variables

Response delay. To operationalize the response delay, we first define the concept *response*. A response is an agent message that follows the previous customer message. To be considered a response, the agent message should be directly preceded by the customer message in the chat conversation (see Figure 2). As a result, not all agent messages will be responses and not all customer messages will have a response. For each pair of customer message and agent response, we find the time that a customer waits to receive an agent response. This value gives us the response delay for the customer message. We account for back-to-back messages as well as canned agent messages or responses in our computations. For back-to-back customer messages, we define the response delay as the gap between the last customer message and the following agent message in our main models. As a robustness check, we also use the gap between the first customer message and the following agent response. For canned messages, we identify three types of

13

messages in the data - greetings, goodbye messages and survey reminders - and exclude them from our computations and analysis. We measure the response delays in seconds.

Problem resolution. Problem resolution is an outcome indicator for the chat conversation that indicates whether or not the customer's problem was successfully resolved by the agent at the end of the conversation. We collect problem resolution information from customer survey responses. This variable is coded in binary, with 1 indicating that the problem was resolved, and 0 otherwise.

Customer satisfaction. Customer satisfaction is measured at the session level and reflects the customer's evaluation of their service encounter (the related chat session) (Tax et al. 1998). We collect customer satisfaction scores from two customer survey questions. The first question is a five-scale item that asks if the customer's expectations were fulfilled at the end of the chat transaction. Customers choose from one of the following options: *fell way short of my expectations*, *fell somewhat short of my expectations*, *met my expectations*, *exceeded my expectations* and *greatly exceeded my expectations*. The second question asks, based on the last chat transaction, what the likelihood is for the customer to promote the service to his or her friends (i.e., net promoter score). This question is measured using a 0-10 scale, with 10 being the highest. The company further "bins" this scale into three levels: 0-6 as *low*, 7-8 as *medium* and 9-10 as *high*.

3.4.2. Independent Variables

Multitasking. Multitasking is described as the act of working on multiple jobs in parallel (Aral et al. 2012). In the live-chat context, we measure multitasking using two metrics. The first metric (henceforth MT_1) captures the substantial workload of the agent by counting the number of unique sessions that the agent sent a message to during a particular response delay. This metric focuses on the active attention of the agent and helps identify multitasking at a very granular level. However, it does not consider existing sessions in which the agent may be inactive. To consider the possibility that in-session but inactive sessions may still have an effect on the cognitive capability of the agent, we define a second metric (henceforth MT_2) based on the number of all sessions open on the agent's service screen during a particular response delay.

For both metrics, we measure multitasking at the message level. A particular benefit of measuring multitasking at the message level is the ability to capture exogenous variations in multitasking values. As discussed earlier, service assignments are made automatically by the system. The live-chat environment is dynamic and customers arrive and leave at random time points. As a result, even during one agent-customer conversation, the agent's level of multitasking values may exogenously change from time to time. We also include quadratic forms of the MT metrics in the message-level analysis to account for potential non-linearity. For session-level models, we average the multitasking values of individual responses in the chat session: Multitasking rate of a chat session ($AvgMT$) is defined as the average of multitasking values of individual responses (*MT*) in that session.

3.4.3. Other Variables for Message-Level Analyses

Skill heterogeneity. Contact centers commonly cross-train their staff in multiple skills to increase agent utilization (Tekin et al. 2009). It is expected that both the count and the heterogeneity of jobs that are being multitasked will have an influence on agent performance and service quality. To account for this issue, we observe heterogeneity among multitasked chat sessions. We use skill keys (problem types defined by the company) to distinguish different skill sets required in different sessions. For each agent response, we calculate the distribution of unique skill keys that are being worked on by that agent at that time using a Herfindahl Index. We construct two such heterogeneity indices to correspond to each of our multitasking metrics. For *MT₁*, the heterogeneity measure is constructed as $SkillHetMT_1 = 1 - \sum_{a=1}^{s}\left(m_{ja}/m_j\right)^2$, where s is the total number of skill keys in the data set, m_{ja} is the number of sessions that have skill key index a and that were sent at least one message during response delay j, and m_j is the number of all sessions that were sent at least one message during response delay j. For *MT₂*, the heterogeneity measure is constructed as: $SkillHetMT_2 = 1 - \sum_{a=1}^{s}\left(o_{ja}/o_j\right)^2$, where s is the total number of skill keys in the data set, o_{ja} is the number of sessions that have skill key index a and that were open during response delay j, and o_j is the number of all open sessions during response delay j. To reflect the heterogeneity of skill sets, we subtract

the summations from 1 in both measures, so that higher values correspond to higher heterogeneity of skills required in the sessions that the agent is simultaneously working on.

Multitasking variation. It is possible that the variability of the multitasking during the session may lead to inconsistency in service delivery, which may have an additional effect on problem resolution and customer satisfaction. To consider this issue, we define a multitasking variation variable based on the standard deviation of multitasking within the chat session.

Typing Effort. The textual length of an agent's message is expected to affect response delays. Naturally, the longer the text that the agent types, the longer it would take to send the message. We consider such effects to be *mechanical* due to the nature of text-based communication. To disentangle the mechanical delays caused solely by typing a text message, we include a message length control in our analyses. We operationalize this variable by counting the total number of words typed by the agent in all the messages that he or she sent during a particular response delay.

Customer message length. The length of the customer message preceding the agent response may affect the delay of the response for two main reasons. First, as the customer's message gets lengthier, it would take longer for the agent to read the entire message. Second, agents may be less enthusiastic about responding to customer messages that are wordy and long-winded. We therefore control for the customer's message length. This variable is defined as the length of the customer message preceding the response for each response in a given session.

Customer slowness. Even though online chats are inherently asynchronous, they are still interactive; a smooth conversation will require the engagement of both the customer and the agent. If the customer is not responsive to agent queries, the agent may be less likely to respond promptly. To account for this issue, we compute the moving average of a customer's message delays (i.e. time delays between agent messages and corresponding customer messages) and include it in the message-level model.

Message order. Not all agent responses are expected to take a similar amount of time. To account for the heterogeneity in response times due to the position of the response, we include an order variable in our

multitasking vs. response delay model. We bin the message order into three groups – first two messages as the greeting phase, last two messages as the feedback phase and in-between messages as the solution phase.

Session key. We use session fixed effects in the multitasking vs. response delay model in order to analyze multitasking at the granularity level of a message. Session dummies help us control for any time invariant characteristics and unobserved heterogeneity within individual chat sessions.

3.4.4. Other Variables for Session-Level Analyses

Queue wait time. Waiting time in queue refers to the amount of the time a customer waits after arriving to the system and before being admitted into service by an agent. It is widely considered to be a major determinant of service quality and customer satisfaction (Durrande-Moreau 1999). Further, waiting time in queue is expected to correlate with multitasking, since company policies dictate that agents switch to multitasking when queues develop in the system. To disentangle these effects, we control for queue wait time in our models. Information regarding waiting times for each customer is extracted from transaction server logs. Waiting times are measured in seconds.

Session duration (i.e., service time). Session duration is the total amount of time a customer spends in service, between the time they were admitted into service and the time they left the system. Similar to queue wait time, this information is obtained from transaction server logs and measured in seconds.

Agent experience. For session-level models, the impact of multitasking may depend on agents' experience levels with the chat sessions. For example, an agent may become more skillful in solving a problem as he or she processes more problems of similar type. To account for this issue, we include an agent experience control in our models. We operationalize agent experience by counting the number of same skill key sessions that the agent has processed during the data collection month prior to each session.

Negative emotion prior. Customers arriving at a chat queue may already be in a particular emotional state. The positivity or negativity of that state will not only affect their interaction with the agent, but also their ultimate evaluation of the entire customer service engagement. We therefore aim to control for such pre-existing emotional state of customers in our analysis. The problem description that customers are

17

prompted by the online chat system to enter is uniquely ideal for this purpose, because it is written before customers are assigned to any particular agent. But since these texts are typically short and therefore not conducive to textual analyses, we combine them with the first three messages customers write after they enter the chat sessions. We then conduct sentiment analysis on those texts to extract the negative emotion of each customer at the beginning of a chat session using the standard text mining software package LIWC (Pennebaker et al. 2001).

Survey response gap. The time difference between the date of a chat session and the survey response date for this chat session could be a factor affecting the customer satisfaction responses in the survey. For example, customers' scoring of an unpleasant transaction may be less harsh as more time passes. To account for this issue, we include a survey response gap variable in our model of customer satisfaction. Survey response gap is computed as the number of days between the chat date and the survey response date. Survey links are sent one day after the chat sessions.

Description Length. In our analyses we control for the length of the initial customer message before s/he is connected to a chat agent.

Agent key. During the data coverage period, there were 1,209 unique service agents who conducted chat sessions with customers. Naturally, these agents have difference experience levels as well as different behaviors. We include agent fixed effects in our models to account for unobserved heterogeneity across individual agents.

Skill key. To control for differences among problem types, we include skill key dummies in our session-level models.

Product key. The contact center is operated by a company that develops tax preparation and filing support software. There are multiple software versions and product types that are offered by this company. These products vary based on their tax filing features and prices. To account for the inherent variation among different products, we include dummy variables for product keys in our session-level models.

Gender. Demographics variables such as gender are commonly used to determine membership in customer segmentation studies (Gupta and Chintagunta 1994). We use gender in our finite mixture model to classify a customer in terms of their tolerance of multitasking. To identify the gender, we extract each customer's first name from chat transcripts using text processing techniques and then run the extracted name through a baby naming dictionary[6] via an automated script.

Contact date and *Contact time.* To account for temporal variations, we include contact date and contact time interval dummies as controls.

Descriptive statistics regarding these variables are given in Tables 3-6. In total, we analyze 14,125 chat sessions and 600,376 individual response messages. We note that message-level and session-level models use their corresponding variables at their own level of granularity. Before estimating the models, we transform control variables into their natural logarithms to reduce skewness[7]. As a robustness test, we also use these variables in the estimations without log transformations. Tables 3 and 4 provide the summary statistics and correlation results for message-level variables, whereas Tables 5 and 6 provide the same information for session-level variables.

<div align="center">INSERT TABLES 3 - 6 HERE</div>

4. Empirical Analyses and Results

4.1. Multitasking and Response Delay

4.1.1. Model Specification and Estimation

To investigate the relationship between multitasking and response delays, we first run OLS estimations with session fixed effects using a panel dataset[8]. The panel dataset includes observations of each individual agent response within the entire chat session and for all chat sessions. Each observation is taken at the time of an agent response; hence, our unit of analysis is an individual message.

---

[6] http://www.gpeters.com/names/baby-names
[7] Except dummy variables such as gender.
[8] Section 5 reports results on robustness tests for the analyses reported here such as matching.

The dependent variable is the response delay (*RDelay*) and the key independent variable is the multitasking amount per response (*MT*), measured in two different forms. We include the quadratic form of the multitasking variable (MT^2) to check for possible non-linearity. We use the skill heterogeneity across the processed sessions (*SkillHetMT*) in the estimation to account for possible cognitive delays due to switching between dissimilar tasks. Other control variables include agent's typing effort (*Typing*), message length of the preceding customer message (*CLength*), the customer's average message delay up to the particular response (*CSlowness*) and the dummy variables for the response order within the session (*Order*).

An econometric concern in our analysis is unobserved heterogeneity. Session fixed effects help us control for time-invariant characteristics of chat sessions such as unique agent characteristics and problem variation as well as unobserved heterogeneity for time of day and date effects. A second concern in our analysis is simultaneity, i.e., multitasking may be caused by delays between messages. However this is an unlikely scenario in our context, because it is the chat system policy that dictates admission control decisions. It is not up to the agents to decide whether they have more than one customer at a time and who that customer is. And since customer arrival and departure times are also largely random, the multitasking variable is not only exogenous but also can vary exogenously from one message to another, even within the same chat conversation. These features allow us to identify the effect of multitasking, measured in the two different ways that we described, on service performance.

4.1.2. Results

Table 7 reports the first estimation results of various specifications. Columns (1) and (2) employ two different multitasking variables independently. Subsequent specifications use both multitasking variables in the same estimation. Columns (4) to (7) use one-way (agent and date levels) and two-way clustered standard errors as well as bootstrapped errors to check the robustness of results.

INSERT TABLE 7 HERE

The results from Table 7 provide support for H1: Multitasking indeed has a positive and statistically significant effect on response delays. The results are consistent across multiple specifications for response

delays. The direction and statistical significance remain consistent for both forms of *MT* variables. However, the degree of impact is different for different definitions of multitasking. Merely having a parallel session in progress (even if the agent is idle in that session) increases response delays by 3 seconds per message[9]. On the other hand, it takes 34 seconds longer for an agent to respond to a particular customer's message, if this agent was actively multitasking with a new parallel session. This amounts to a slowdown of 60% in agent responses. From the customer's perspective, an interpretation of this finding is that each customer spends 4 extra minutes[10] in service when the agent is handling one additional customer in parallel.

We are also interested in the marginal effects of increasing multitasking levels on response delays. The coefficients on the multitasking squared terms (for both MT measures) are positive and significant, implying a convex relationship. This means that multitasking an additional customer has a greater negative impact on response delays for agents who concurrently process more jobs, as opposed to agents who concurrently process fewer jobs. This finding provides support for the second argument of H1.

Before we conclude this section, we turn to the findings on the message order. Our base case for binned order variables is Bin=1, i.e., the beginning phase of the chat transaction. We find that it takes about 5.6 seconds longer to respond to message in the solution phase than in the base case. On the other hand, compared to the base case again, messages in the closure phase take about 7.3 seconds quicker to respond. This finding highlights the inherent response delay differences between different stages of the conversation.

4.2. Multitasking and Problem Resolution

4.2.1. Model Specification and Estimation

To test the hypothesis that multitasking decreases service agents' ability to successfully resolve problems, we estimate a logit model on the session-level data. The outcome variable is 1 if a customer's problem is

---

[9] Being idle in a chat session is a rare and unlikely situation for agents. Therefore, this value is expected to be a conservative estimate for the true impact of multitasking.

[10] This value is for the average session in the dataset (which has 7 agent responses) and assumes that the agent is equally distributing his/her workload among two parallel sessions.

resolved at the end of that particular chat session, and 0 otherwise. We collect problem resolution information from customer survey responses. Our main independent variable of interest is the average multitasking level across the session (*AvgMT*). We include the average skill heterogeneity across the session (*AvgSkillHetMT*) to account for possible cognitive effects caused by working on dissimilar tasks simultaneously. We also include the standard deviation of multitasking across the session (*sdMT*) to account for the variability of multitasking possibly affecting the problem resolution. The control variables include waiting time in queue (*QueueWait*), session duration (*Duration*), the experience level of the agent with the particular skill for the session (*AExperience*), negative emotions of the customer at the beginning of the session (*NegEmotion*), the brevity of the description of the problem at the beginning of the session (*DescLength*), the total number of words that the agent typed (*TotalTyping*), average customer message length (*AvgCLength*), response delay (*AvgCSlowness*), and the number of days between the contact date and the survey response date (*SurveyGap*). These values are computed using text processing techniques on chat transcripts and server logs. Similar to the message-level model, we use two measures of *MT* in our estimations. We include agent fixed effects in the model to control for unobserved agent-specific heterogeneity. We also add skill key, product key, contact date and contact time interval fixed effects into the model to control for other forms of heterogeneity. Skill key and product key fixed effects aim to address inherent differences in problem complexity and product differences. Contact date and contact time interval fixed effects are used to control for time-specific trends and differences.

Another empirical issue to consider in the analysis is self-selection bias. Customer satisfaction surveys are especially prone to self-selection bias because of possible systematic motivational differences between those who choose to respond to surveys and those who do not. As stated earlier, 16% of customers in our entire data set completed the customer satisfaction surveys. While this response rate is on par with the

industry average of 10-15% for external surveys[11], it could still indicate a selectivity bias. We address this potential bias by using the two-step Heckman correction method (Heckman 1979). First, we estimate the likelihood of a customer responding to the survey using Probit regression. The Probit model uses session duration, queue wait time, agent experience, prior negative emotion and description length as predictor variables and constructs the inverse Mills ratio (λ). λ accounts for the fact that self-reported resolution information is only observed when customers respond to the surveys. Then, we include λ in the estimation of our main model.

4.2.2. Results

Table 8 reports the results for multitasking vs. problem resolution analysis for various specifications. The first two columns provide the estimations for MT_1 and MT_2 individually, whereas Column (3) uses both multitasking variables in the same estimation. Columns (4) and (5) provide the results with clustered standard errors. We note that due to the inclusion of agent fixed effects and some of the agents' indifference in terms of the dependent variable, 618 observations are removed from the logit model estimation.

INSERT TABLE 8 HERE

Three primary results emerge based on the estimation of this model. First, among all the specifications, the coefficient of $AvgMT_1$ is statistically significant and negative, indicating that working on a parallel session is negatively associated with problem resolution, providing support for H2. The coefficient of $AvgMT_1$ (Columns 3-5) indicates that if an agent actively communicates with this additional customer throughout the existing session, then the estimated odds of resolving the existing problem decreases by approximately 51%. This result is robust to unobserved heterogeneity and self-selection bias, and it is consistent across different standard error structures. Second, we do not observe the same degree of impact for the $AvgMT_2$, since its coefficient (when estimated together with $AvgMT_1$) is substantially smaller and

---

[11] See, for example, https://www.surveygizmo.com/survey-blog/survey-response-rates/.

insignificant. This finding supports the argument that there is a difference between two different definitions of multitasking: 1) being assigned an additional task and 2) actively working on an additional task. Problem resolution capability of agents diminishes when the agents actively switch between different tasks throughout their work periods. Considering the fact that multitasking policies inherently promote task-switching behavior[12], this situation appears to create an interesting conundrum for agents – whether to follow the policy for better utilization of their time or ignore the policy in return for more successfully resolving their assigned tasks. Finally, our finding about the decline in problem resolution due to multitasking not only raises concerns on the quality of service under multitasking regimes, but also challenges the argument that favors multitasking for its presumed productivity benefits. For many service organizations with repeat customers, such presumed productivity gains may cease to materialize in the long run, because customers whose problems have not been resolved may return later (Mehrotha et al. 2012), either to the same customer service channel or a different one (e.g., phone or in-person).

4.3. Multitasking and Customer Satisfaction

4.3.1. Model Specification and Estimation

To study the relationship between multitasking and customer satisfaction, we conduct a path analysis for the model shown in Figure 3.

<div align="center">INSERT FIGURE 3 HERE</div>

This analysis aims to identify the possible direct effect of multitasking on customer satisfaction, as well as possible indirect effects via response delay and problem resolution paths. For this purpose, we fit three OLS regression models to the session-level data and estimate the path coefficients using the results from the models. In these models, our main variable of interest is the average multitasking level across the session (*AvgMT*), represented in two forms. Our dependent variables are average response delay within the

---

[12] Because the presumed productivity benefits of multitasking can only be realized when agents switch between different tasks.

session (*AvgRDelay*), problem resolution (*Resolution*) and customer satisfaction (*Satisfaction*). We measure *Satisfaction* using two items. The first item is a survey question that asks if the expectations of the customer were fulfilled at the end of the chat session. The second item is a survey question that asks, based on the last chat transaction, the likelihood that the customer will promote the service to friends and relatives.

The three regression models in the path analysis are all at the session level, therefore we use the same procedure discussed in Section 4.2.1 to address econometric concerns. Specifically, we include control variables and agent, skill key, product key, contact date and contact time interval fixed effects in the model to control for different forms of heterogeneity. In addition, we address the self-selection bias using the Heckman correction method.

4.3.2. Results

As discussed earlier, we use two measures for customer satisfaction, which are expectation fulfillment and net promoter scores. Tables 9 and 10 present the results for analyses with each measure, respectively.

<div align="center">INSERT TABLES 9 - 10 HERE</div>

Panel A of these tables shows the regression results for the direct paths from multitasking to response delay and problem resolution, as well as the direct effect on customer satisfaction. $AvgMT_1$ is significant in the first two regressions, and the direction of the coefficient is consistent with our analyses at the message level (Section 4.1) and the logit model (Section 4.2). On the other hand, $AvgMT_1$ is not a significant predictor for customer satisfaction when response delay and problem resolution are controlled for. The results are consistent whether $AvgMT_1$ is used independently or jointly with $AvgMT_2$. This finding points to the indirect effects of multitasking on customer satisfaction, which are calculated in Panel B of the tables. In these panels, we break down our estimations for the effect of multitasking on customer satisfaction into: 1) indirect effect via the response delay path, 2) indirect effect via the resolution path, 3) direct effect and 4) total effect. We compute the indirect effects on the outcome variables by taking the product of coefficients from the corresponding columns of the tables. We find that the total effects of $AvgMT_1$ on expectation fulfillment and net promoter scores are significant and negative. Yet, these total effects are primarily driven

by indirect effects through response delays and problem resolution paths. This finding provides support for H3a and H3b for the MT_1 variable. In other words, while customers cannot observe the efforts of agents directly, they still feel the side effects of multitasking. When a service agent actively engages in communication with multiple customers, the service he or she provides is less likely to fulfill these customers' expectations due to increasing in-service delays and decreasing problem resolution. Further, these customers would be less likely to promote the service to their friends and relatives.

5. Robustness and Additional Analyses

5.1. Robustness

5.1.1. *Matching Analysis at the Session Level*

As discussed earlier, a useful characteristic of our context is the exogenous nature of multitasking assignments. In this contact center setting, customers are served on a first-come-first-serve basis and multitasking decisions are made by a contact distribution system based on the length of the queues. First, to validate that this admissions policy was indeed implemented in our data set, we search the data for records that were exceptions to the policy guidelines. In particular, we search for cases, where a customer was *jumped over* – i.e., got admitted later than a same problem type customer, who came after than him/her. We find an extremely small number of cases (less than 0.1% of all records) that contradict the policy. This finding supports the argument regarding the exogenous nature of customer admissions. Next, to further test the robustness of the session-level results obtained from our setup, we employ a propensity score matching (PSM) approach. This approach helps overcome the potential concerns regarding the queue lengths affected by individual agent characteristics such as speed or capability.

To conduct the PSM analysis, we first estimate a logit model to identify whether a customer would be assigned to a multitasking agent during his or her session (i.e., treated or not). Since our multitasking variables are continuous by definition, we use different thresholds (between 1.3 and 1.8 with 0.1 increments) to define what constitutes a multitasking treatment in a session. For the treatment model, we use *queue wait time*, *agent experience*, *agent key*, *problem skill key*, *contact date* and *contact time interval*

26

as predictor variables. For each observation that was actually treated, we find a non-treated observation that is highly similar in terms of its treatment model score, using nearest neighbor matching.

Tables 11 and 12 provide the results from propensity score matching for each of the multitasking variables (MT_1 and MT_2). From Table 11, we can see that both multitasking definitions have a significant and positive effect on average response delays, regardless of the threshold value. This result is consistent with the regression estimations in Panel A in Tables 9 and 10 (first three columns) and it also complements the findings in the message-level analysis in Table 7. From Table 12, we can see that only $AvgMT_1$ has a consistent negative and significant (for certain thresholds) effect on problem resolution. Again, this finding is consistent with our logit model in Table 8 and the regression models in Tables 9 and 10 (columns 4-6).

<div align="center">INSERT TABLES 11 - 12 HERE</div>

One common concern for propensity score matching is the existence of unobservables that may be driving the difference between the treatment (i.e., multitasking) and comparison (i.e., non-multitasking) groups. In order to investigate how sensitive our matching estimates are to the possible existence of an unobserved confounding variable, we conduct sensitivity analysis (Rosenbaum 2002). We find that in order to attribute the effect to an unobserved confounder, this unobserved confounder would need to produce, on average, a threefold increase in the odds of multitasking. This is much higher than typical thresholds used in prior literature (e.g., Sun and Zhu, 2013, p. 2327), suggesting that unobservable information is unlikely to drive our findings.

5.1.2. *Matching Analysis at the Message Level*

A different concern in our analyses is that at the message level, a service agent's decision on whom to respond to (in the case of handling multiple simultaneous sessions) is based on the agent's choice, therefore MT_1 may be endogenous. First, we note that this is an unlikely event in our context and the response preferences are mainly driven by an external factor that is exogenous to agents. The reason for this exogeneity lies within the nature of multitasking policies. The main premise of multitasking is to improve agent productivity by switching to a different customer whenever there is a downtime with the other

27

customer(s). Productivity benefits is the major reason that multitasking policies are implemented by firms and, therefore, multitasking behavior is primarily driven by the downtimes caused by customers rather than by arbitrary agent decisions. In other words, service agents are more likely to follow the policy for its intended purpose, rather than making arbitrary decisions on whom to serve by themselves.

Nevertheless, to further alleviate this concern, we perform matching analysis at the message level. For this analysis, we apply 1-to-1 matching and match each message in which the agent had performed multitasking (i.e. sent at least one message to a non-focal customer) to another message in a different session in which the agent had strictly worked with a single customer. For the matching procedure, we search for similar characteristics including the same agent, same skill, same product type, same order bin and similar agent and customer message lengths[13]. If there are multiple matches for a particular message, we randomly pick one match and discard the rest. Next, we obtain the response delays for each message in the two groups (multitasked vs. not-multitasked), and run a paired t-test to see if the response delay means across the two groups differ. Table 13 provides the results from this analysis.

INSERT TABLE 13 HERE

We find that paired t-test results are fully consistent with our earlier findings – i.e., there is a statistically significant difference ($t = 43.503, p = 0.000$) in response delays between multitasked ($\mu = 69.026$) and non-multitasked ($\mu = 52.876$) groups.

We run two further robustness tests and find that the results are still robust: 1) rather than using a random pick, we average the response delay from all matches; and 2) rather than using just a t-test, we use

---

[13] This matching procedure is, in a sense, stricter than propensity score matching, because we require matches to be identical on many dimensions. For robustness we also apply the typical propensity score matching procedure. The ATET (average treatment effect on the treated) is consistent with our main findings for both MT_1 and MT_2. Furthermore, the propensity matching method allows us to test the sensitivity of the results to unobservables. We found that for the unobserved confounder to be driving our results, this unobserved factor needs to produce a 4.8 times (for MT_1) or 1.8 times (for MT_2) increase in the odds of multitasking, which is again much higher than the thresholds reported in the literature (e.g., Sun and Zhu 2013).

a multivariate regression that incorporates agent, skill key, product key, contact date, contact time, and order bin fixed effects. For brevity we do not report those results here.

5.1.3. *Non-logged Control Variables*

Our main models employ natural log transformations of control variables to reduce the skewness of data. To ensure that this process doesn't bias the findings, we estimate the response delay and problem resolution models using the original scales of these variables. To conserve space, we provide the full results in Appendix A. The results are highly consistent with those provided in Section 4.

5.1.4. *Alternate Response Delay Measure for Back-to-Back Messages*

An empirical challenge in measuring response delay is the handling of back-to-back messages, since some agent responses were preceded by multiple consecutive customer messages. We previously computed response delays using the final message in customers' messages. This idea is illustrated as Point (1) in Figure 4.

INSERT FIGURE 4 HERE

This approach is expected to lead to a more conservative estimate of the impact of multitasking, and that is the reason that we use it in the main models. To ensure the robustness of our findings, we measure response delays using the first message in the message sequence (illustrated as Point (2) in Figure 4) as a robustness test. We also re-compute the multitasking measurement values accordingly. To conserve space, we provide the full results in Appendix B. These results are consistent with the original analysis and support the finding that multitasking has a positive and statistically significant effect on response delays.

5.2. Acting on Our Findings: A Segmentation Model for a Customer-Specific Routing Policy

Our analyses so far focus on the cause-effect relationships between multitasking and the dependent variables of interest. These findings can first and foremost inform managerial decisions in terms of whether to multitask agents, and to what extent agents should be multitasking. But given budgetary constraints, companies may not always be able to avoid multitasking, so the natural question is: *Can we selectively let*

agents multitask in some situations and not multitask in others? Can our analyses be extended to provide design guidelines for the customer assignment systems to preserve customer satisfaction while increasing the throughput? A full treatment of this question will be beyond the scope of this paper, thus here we focus on one angle: customer heterogeneity in terms of their reactions to, or rather tolerance of, agent multitasking. If we can use some information about customers to classify them into sensitive and non-sensitive clusters, then the system can assign less sensitive customers to multitasking agents. Doing so would help us leverage the productivity benefits of multitasking (in terms of more efficient use of agents' times), while reducing its negative impact on customer satisfaction.

We propose a finite mixture approach that takes such possible customer heterogeneity into account. Finite mixture approaches have been commonly used in the marketing domain for customer segmentation (Wedel and Kamakura 2000) and are being increasingly adopted in IS research (Bapna et al. 2011). Following common practice, we assume that customers in our dataset can be classified into *S* discrete segments (i.e., classes). Customers within the same segment are assumed to be homogeneous in terms of their satisfaction behavior based on session-specific variables. However, sensitivities towards these variables may differ across different segments. We define the probability structure for this model as follows:

Let $P_i(k|s)$ denote the probability that at the end of chat session *i,* the customer will have the satisfaction score of *k* conditional on the customer belonging to segment *s*. Using a multinomial logistic regression model, this probability can be represented as:

$$P_i(k|s) = \frac{exp(\mathbf{B}_s\mathbf{X}_{ik})}{\sum_{l=1}^{K} exp(\mathbf{B}_s\mathbf{X}_{il})} \tag{1}$$

Where $\mathbf{X}_{ik}$ is the column vector of explanatory variables that can help explain the satisfaction behavior and $\mathbf{B}_s$ is the row vector of coefficients associated with $\mathbf{X}_{ik}$. Further, let P_{is} denote the probability that a customer will belong to segment *s*. P_{is} depends on a vector of customer-specific (i.e., concomitant) variables $\mathbf{D}_i$, which can be represented as:

$$P_{is} = \frac{exp(\boldsymbol{\gamma}_s \mathbf{D}_i)}{\sum_{s=1}^{S} exp(\boldsymbol{\gamma}_s \mathbf{D}_i)} \tag{2}$$

Where $\boldsymbol{\gamma}_s$ $(s = 1,2,\dots,S)$ is the vector of coefficients that represents the effects of customer-specific variables on the probability of segment membership. Combining these two equations, the probability of an arbitrary session i having satisfaction level k is:

$$P_i(k) = \sum_{s=1}^{S} P_{is} \times P_i(k|s) = \sum_{s=1}^{S} \frac{exp(\boldsymbol{\gamma}_s \mathbf{D}_i)}{\sum_{s=1}^{S} exp(\boldsymbol{\gamma}_s \mathbf{D}_i)} \times \frac{exp(\mathbf{B}_s \mathbf{X}_{ik})}{\sum_{l=1}^{K} exp(\mathbf{B}_s \mathbf{X}_{il})} \tag{3}$$

Estimation of the unknown coefficients $\mathbf{B}_s$ and $\boldsymbol{\gamma}_s$ can be conducted using maximum likelihood method. Once these estimates are obtained, we can assign each customer to a segment by finding the segment for which the customer has the highest posterior probability.

To demonstrate the application of this model, we apply it within the context of multitasking – customer satisfaction analysis. We assume that there are two segments of customers who have different levels of sensitivity towards in-service delays[14]. At the time of their arrival to the system, we have limited information about the customers. Specifically, we can use the following variables to classify the customers into one of the two segments: gender (*Gender*), problem domain (*Skill domain*), product type (*Product key*) and hourly time interval of the contact (*Contact time*). The values of these variables are recurrent and easy to obtain, making them feasible for use in the segmentation process. A customer satisfaction estimation can then be conducted for each segment.

Table 14 presents the finite mixture estimation results with *Satisfaction₁* as the dependent variable and *AvgMT₁* as one of the predictor variables. We provide results for both 1-class and 2-class models to compare different fits. Four different model choice criteria (AIC, BIC, CAIC and R-squared) collectively suggest that a 2-class model is a better fit than a 1-class model for this dataset. For the 2-class model, the *p*-value

---

[14] While there may be different values for S that would fit the data better, we pick $S = 2$ since our main goal is to demonstrate the applicability of finite mixture models for deriving actionable strategies. A two segment model could guide the contact center to implement a simple two-class routing policy, i.e., use different routing rules for customers who are more (less) sensitive to multitasking. Models with more segments will also require more customer covariates.

for the Wald statistics of $AvgMT_1$ is less than 0.001, indicating that the overall effect of multitasking is highly significant. The coefficient of this effect is negative and statistically significant for both segments, which is consistent with our previous findings. On the other hand, at the individual-segment level, we find the coefficient of $AvgMT_1$ to be approximately 2 times larger than that of Segment 2, indicating a much stronger negative impact of multitasking for customers that are members of Segment 1. A separate Wald statistic shows that the between-segment difference in terms of $AvgMT_1$ indeed exists and is significant ($p < 0.1$). In terms of identifying the segments, we find the coefficients of gender (*Gender*), problem domain (*SkillDomain*) and product type (*ProductID*) to be significant, meaning that these variables can be used for customer segment prediction.

INSERT TABLE 14 HERE

In summary, these results suggest that multitasking has a negative overall effect on customer satisfaction; however not all customers are affected to the same degree. By utilizing information from the time of customer arrival to the system, contact centers can segment customers into two groups with different multitasking sensitivity levels. This segmentation scheme can further be integrated into a routing policy that can make decisions on which customers should be multitasked first when multiple customers are waiting for service.

6. Conclusion

In this research, we studied the negative consequences of multitasking in service organizations. We utilized a diverse set of operational data (including contact transcripts, server logs and customer survey responses) from an online chat contact center to investigate the effects of multitasking on problem resolution rates, response delays and customer satisfaction. Our results are based on two complementary metrics of multitasking: 1) whether an agent was nominally assigned an additional customer and 2) whether he/she was actively working with that additional customer. We also expanded our analyses to account for customer heterogeneity and identified tailored routing strategies based on customer segmentation. Prior research has mainly followed analytical and experimental approaches to examine the impact of multitasking on

productivity (i.e., throughput) related measures. To our knowledge, this is one of the very few studies that investigates the multitasking phenomenon from a service quality perspective using observational data.

Our findings have several implications for research and practice. Previous research has shown that under certain conditions, it is possible to increase productivity via multitasking (Aral et al. 2012; O'Leary et al. 2011). We found that increased productivity may come at the expense of diminishing quality of service and customer satisfaction. A managerial implication of this finding is that focusing solely on the productivity measures may not be the best benchmark for firms. Service organizations need to take a holistic view of their service goals, from both quality of service and productivity perspectives.

An interesting aspect of our context is that in live-chat communication, customers cannot directly observe the task-processing behavior of agents and they may not be aware of the on-going multitasking activity. Yet, they still feel the side effects of multitasking inadvertently. First, customers experience longer in-service delays in the presence of multitasking. This result holds even if the agent focuses on only one customer during service. However, the effect is much greater when the agent routinely switches between multiple tasks or when the number of multitasked jobs is increased (i.e., non-linear impact). Second, customers are less likely to get their problems resolved if their agents are communicating with different customers in parallel. We found that the same result does not hold when a service agent gets assigned a new task, but he or she does not actively work on this task. This is an interesting finding, which brings a new perspective on the literature about the impact of discretionary (i.e., voluntary) task switching (Madjar and Shalley 2008; Payne et al. 2007). Our results further suggested that the presumed productivity gains of multitasking may cease to materialize in the long run because customers who have received unsatisfactory service may return to the system and create future demands. Overall, we demonstrated that multitasking has an indirect, but significant, negative impact on service quality and customer satisfaction through in-service delay and problem resolution paths.

Finally, our study contributes to the extensive literature on the well-known control problem in contact centers known as the call routing problem (Aksin et al. 2007) from a unique angle. Existing work in this

area has tackled this problem mostly from a theoretical perspective using stylized analytical models. We

showed that explanatory and predictive (i.e., segmentation) models built on observational data can be

simultaneously used to develop simple but actionable routing strategies. We believe that such routing

strategies are widely applicable and can have significant practical implications for contact centers.

7. References

Adler, R., and R. Benbunan-Fich. 2012. "Juggling on a high wire: Multitasking effects on performance." *International Journal of Human-Computer Studies* 70 (2): 156-168.

Aksin, Zeynep, Mor Armory, and Vijay Mehrotra. 2007. "The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research." *Productions and Operations Management* 665-688.

Anderson, Eugene W., Claes Fornell, and Roland T. Rust. 1997. "Customer Satisfaction, Productivity, and Profitability: Differences between Goods and Services." *Marketing Science* 16 (2): 129-145.

Aral, Sinan, Erik Brynjolfsson, and Marshall Van Alstyne. 2012. "Information, Technology and Information Worker Productivity." *Information Systems Research* 23 (3): 849-867.

Bailey, B. P., and J. A. Konstan. 2006. "On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state." *Computers in Human Behavior* 685-708.

Bapna, R., P. Goes, K. K. Wei, and Z. Zhang. 2011. "A Finite Mixture Logit Model to Segment and Predict Electronic Payments System Adoption." *Information Systems Research* 22 (1): 118-133.

Bavafa, H., L.M. Hitt, and C. Terwiesch. 2017. "The Impact of e-Visits on Visit Frequencies and Patient Health: Evidence from Primary Care." URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2363705. Retrieved on January 27, 2017.

Brady, Michael K., and Joseph J. Cronin. 2001. "Some New Thoughts on Conceptualizing Perceived Service Quality: A Hierarchical Approach." *Journal of Marketing* 65 (3): 34-49.

Cameron, Ann-Frances, and Jane Webster. 2013. "Multicommunicating: Juggling Multiple Conversations in the Workplace." *Information Systems Research* 24 (2): 352-371.

Cameron, Ann-Frances, and Jane Webster. 2011. "Relational Outcomes of Multicommunicating: Integrating Incivility and Social Exchange Perspectives." *Organization Science* 22 (3): 754-771.

Campello, F., A. Ingolfsson, and Robert A. Shumsky. 2016. "Queueing Models of Case Managers." *Management Science*. Published online in Articles in Advance.

Charron, S., and E. Koechlin. 2010. "Divided Representation of Concurrent Goals in the Human Frontal Lobes." *Science* 328 (5976): 360-363.

Chebat, Jean-Charles, and Pierre Filiatrault. 1993. "The Impact of Waiting in Line on Consumers." *International Journal of Bank Marketing* 11 (2): 35-40.

Coviello, D., A. Ichino, and N. Persico. 2014. "Time Allocation and Task Juggling." *American Economic Review* 104 (2): 609-623.

Czerwinski, M., E. Horvitz, and S. Wilhite. 2004. "A Diary Study of Task Switching and Interruptions." *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI).*

Demers, Alan, Srinivasan Keshav, and Scott Shenkar. 1989. "Analysis and simulation of a fair queueing algorithm." *Proceedings of SIGCOMM'89.* 1-12.

Diaz, Ana B. C., and Francisco J. M. Ruiz. 2002. "The consumer's reaction to delays in service." *International Journal of Service Industry Management* 13 (2): 118-140.

Dube-Rioux, Laurette, Bernd H. Schmitt, and France Leclerc. "Consumers' Reactions to Waiting: When Delays Affect the Perception of Service Quality." *Advances in Consumer Research* 16 (1989): 59-63.

Durrande-Moreau, Agnes. 1999. "Waiting for service: ten years of empirical research." *International Journal of Service Industry Management* 10 (2): 171-194.

Fornell, Claes, Michael D. Johnson, Eugene W. Anderson, Jaesung Cha, and Barbara E. Bryant. 1996. "The American Customer Satisfaction Index: Nature, Purpose, and Findings." *Journal of Marketing* 60: 7-18.

Gupta, S., and P. K. Chintagunta. 1994. "On Using Demographic Variables to Determine Segment Membership in Logit Mixture Model." *Journal of Marketing Research* 31 (1): 128-136.

Hall, N. G., J. Y.-T. Leung, and C.-L Li. 2015. "The Effects of Multitasking on Operations Scheduling." *Production and Operations Management* Article in press.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153-161.

Human Capital Management Institute. 2010. "Managing an Organization's Biggest Cost: The Workforce." White Paper.

KC, D. S. 2013. "Does Multitasking Improve Performance? Evidence from the Emergency Department." *Manufacturing & Service Operations Management* 16 (2): 168-183.

Luo, J., and J. Zhang. 2013. "Staffing and Control of Instant Messaging Contact Centers." *Operations Research* 61 (2): 328-343.

Madjar, N., and C. E. Shalley. 2008. "Multiple Tasks' and Multiple Goals' Effect on Creativity: Forced Incubation or Just a Distraction?" *Journal of Management* 34 (4): 786-805.

Mark, Gloria, Daniela Gudith, and Ulrich Klocke. 2008. "The Cost of Interrupted Work: More Speed and Stress." *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI).*

Mehrotha, V., K. Ross, F. Ryder, and Y-P Zhou. 2012. "Routing to Manage Resolution and Waiting Time in Call Centers with Heterogeneous Servers." *Manuf. & Serv. Oper. Management* 14 (1): 66-81.

O'Leary, Michael B., Mark Mortensen, and Anita W. Woolley. 2011. "Multiple Team Membership: A Theoretical Model of its Effects on Productivity and Learning for Individuals and Teams." *Academy of Management Review* 36 (3): 461-478.

Payne, S. J., G. B. Duggan, and H. Neth. 2007. "Discretionary task interleaving: heuristics for time allocation in cognitive foraging." *Journal of Experimental Psychology* 136 (3): 370-388.

Pennebaker, J. W., R. J. Francis, and M. E. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC).* Mahwah, NJ: Erlbaum.

Reinsch, N. Lamar, Jeanine W. Turner, and Catherine H. Tinsley. 2008. "Multicommunicating: A Practice Whose Time Has Come?" *Academy of Management Review* 33 (2): 391-403.

Rosenbaum, P. R. 2002. *Observational Studies.* NY: Springer.

Sun, M., Zhu, F. 2013. Ad Revenue and Content Commercialization: Evidence from Blogs. *Management Science.* 59(10) 2314-2331.

Szymanski, David M., and David H. Henard. 2001. "Customer Satisfaction: A Meta-Analysis of the Empirical Evidence." *Journal of the Academy of Marketing Science* 29 (1): 16-35.

Tan, T. F., and S. Netessine. 2014. "When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity." *Management Science* 60 (6): 1574-1593.

Tan, Yong, Vijay S. Mookerjee, and Kamran Moinzadeh. 2005. "Optimal processing policies for an e-commerce web server." *INFORMS Journal on Computing* 17: 99-110.

Tax, Stephen S., Stephen W. Brown, and Murali Chandrashekaran. 1998. "Customer Evaluations of Service Complaint Experiences: Implications for Relationship." *Jour. of Marketing* 62 (2): 60-76.

Taylor, Shirley. 1994. "Waiting for Service: The Relationship between Delays and Evaluations of Service." *Journal of Marketing* 58 (2): 56-69.

Tekin, Eylem, W. J. Hopp, and M. P. van Oyen. 2009. "Pooling strategies for call center agent cross-training." *IIE Transactions* 41: 546-561.

TELUS International. 2011. *Best Practices: Online Chat Sales.* TELUS.

Tezcan, T., and J. Zhang. 2014. "Routing and Staffing in Customer Service Chat Systems with Impatient Customers." *Operations Research* 62 (4): 943-956.

Tom, Gail, and Scott Lucey. 1995. "Waiting time delays and customer satisfaction in supermarkets." *Journal of Services Marketing* 9 (5): 20-29.

Wedel, M., and W. A. Kamakura. 2000. *Market Segmentation: Conceptual and Methodological Foundations.* 2nd. Boston: Kluwer Academic Publishers.

Zeithaml, Valarie A., Leonard L. Berry, and A. Parasuraman. 1996. "The Behavioral Consequences of Service Quality." *Journal of Marketing* 60 (2): 31-46.

FIGURES
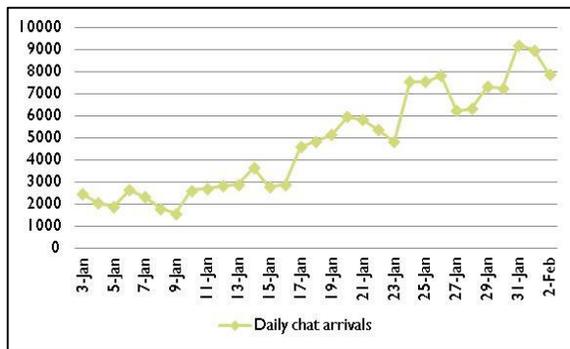
Figure 1a: Daily Chat Arrivals
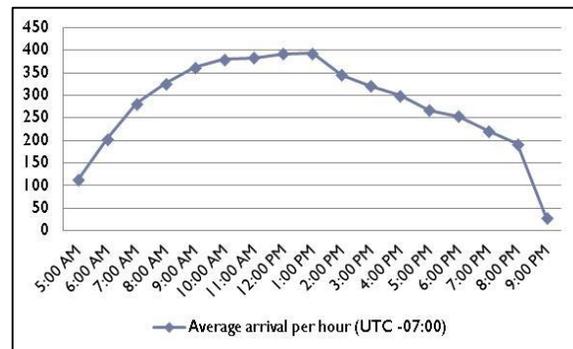
Figure 1b: Hourly Chat Arrivals (MST)
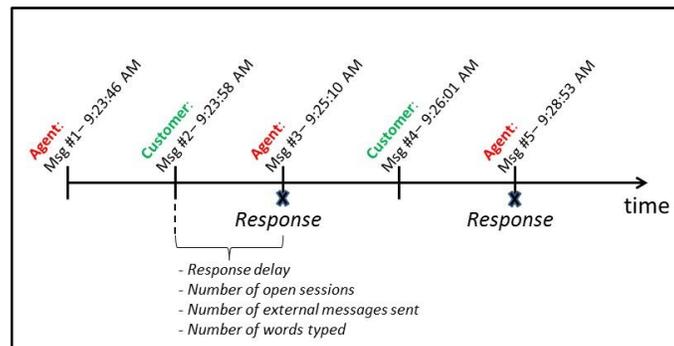
Figure 2: Illustrative Agent Responses in a Conversation Timeline
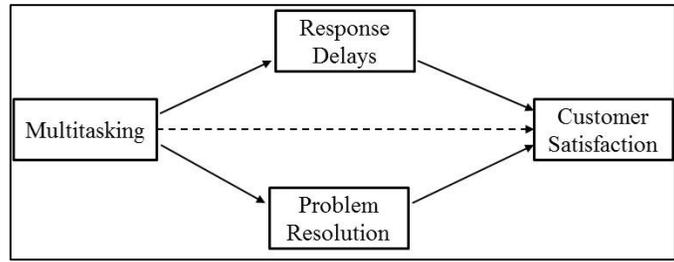
Figure 3: Indirect Relationship between Multitasking and Customer Satisfaction
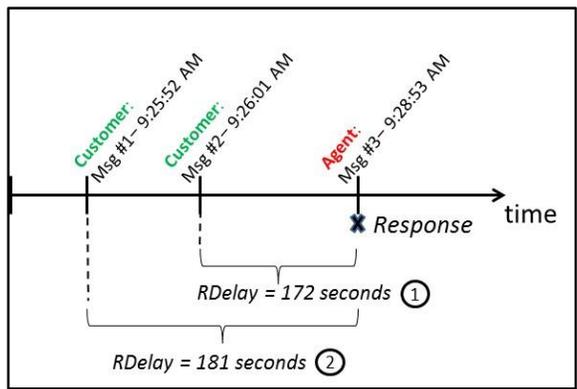
Figure 4: Two Approaches to Handle Back-to-Back Messages

TABLES

Table 1: Data Sources

	Transaction Server Logs	Chat Transcripts	Survey Response
Chat Session ID	✔	✔	✔
System features	✔		
Session features	✔	✔	
Problem description		✔	
Multitasking info.	✔	✔	
Customer satisfaction			✔
Problem resolution			✔

Table 2: Summary Statistics of the Contact Center *(after data cleaning)*

Total number of chat sessions:	86,159
Total number of surveys returned:	14,125
Total number of agents:	1,209
Total number of distinct skills:	20
Total number of agent responses:	600,376
Avg. number of skills that an agent has:	6
Avg. number of agent responses in a session:	7
Avg. number of messages exchanged in a session:	17
Avg. number of words typed per agent response:	16

Table 3: Descriptive Statistics of Message-Level Variables

	Variable	Obs.	Mean	Std. Dev.	Min	Median	Max
(a)	$RDelay$	600,376	56.47	60.04	1.00	38.00	597.00
(b)	MT_1	600,376	1.30	0.51	1.00	1.00	4.00
(c)	$SkillHetMT_1$	600,376	0.10	0.20	0.00	0.00	0.75
(d)	MT_2	600,376	1.91	0.66	1.00	2.00	4.00
(e)	$SkillHetMT_2$	600,376	0.28	0.26	0.00	0.44	0.75
(f)	$Typing$	600,376	25.37	27.14	1.00	17.00	883.00
(g)	$log(Typing)$	600,376	2.76	1.06	0.00	2.83	6.78
(h)	$CLength$	600,376	9.42	10.60	1.00	6.00	110.00
(i)	$log(CLength)$	600,376	1.91	0.94	0.00	1.95	4.71
(j)	$CSlowness$	600,376	37.18	36.06	1	26	259
(k)	$log(CSlowness)$	600,376	3.17	1.05	0	3.26	5.56

Table 4: Correlation Matrix for Message-Level Variables

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
(a)	1.00										
(b)	0.39	1.00									
(c)	0.28	0.79	1.00								
(d)	0.14	0.43	0.37	1.00							
(e)	0.09	0.27	0.47	0.69	1.00						
(f)	0.45	0.62	0.46	0.26	0.16	1.00					
(g)	0.39	0.58	0.46	0.23	0.14	0.79	1.00				
(h)	0.05	0.01	0.02	0.04	0.04	0.03	0.05	1.00			
(i)	-0.01	-0.02	0.00	0.04	0.04	0.00	0.01	0.88	1.00		
(j)	0.05	0.02	0.03	0.05	0.04	0.02	0.01	0.34	0.33	1.00	
(k)	0.06	0.03	0.03	0.04	0.03	0.03	0.02	0.29	0.32	0.80	1.00

Table 5: Descriptive Statistics of Session-Level Variables

Variable	Obs.	Mean (Freq.)	Std. Dev.	Min	Median	Max
(l) $AvgMT_1$	14,125	1.32	0.32	1.00	1.25	4.00
(m) $AvgSkillHetMT_1$	14,125	0.10	0.13	0.00	0.06	0.67
(n) $sdMT_1$	14,125	0.36	0.28	0.00	0.45	2.12
(o) $AvgMT_2$	14,125	1.94	0.60	1.00	2.00	4.00
(p) $AvgSkillHetMT_2$	14,125	0.30	0.24	0.00	0.36	0.75
(q) $sdMT_2$	14,125	0.19	0.25	0.00	0.00	1.41
(r) $log(QueueWait)$	14,125	3.77	2.57	0.00	4.16	7.88
(s) $log(Duration)$	14,125	6.91	0.73	3.78	6.90	8.82
(t) $log(AExperience)$	14,125	2.88	1.66	0.00	2.83	6.94
(u) $log(NegEmotion)$	14,125	0.28	0.46	0.00	0.00	2.48
(v) $log(DescLength)$	14,125	3.93	0.63	0.69	3.99	6.45
(w) $log(TotalTyping)$	14,125	5.01	0.81	0.00	5.13	7.30
(x) $log(AvgCLength)$	14,125	2.30	0.55	0.69	2.32	4.30
(y) $log(AvgCSlowness)$	14,125	3.62	0.49	0.00	3.62	5.55
(z) $log(SurveyGap)$	14,125	0.95	0.44	0.69	0.69	5.06
(aa) $Resolution$	14,125	0.63	0.48	0.00	1.00	1.00
(bb) $Satisfaction_1$	14,021	2.99	1.40	1.00	3.00	5.00
(cc) $Satisfaction_2$	14,069	2.24	0.88	1.00	3.00	3.00

Table 6: Correlation Matrix for Session-Level Variables

	(l)	(m)	(n)	(o)	(p)	(q)	(r)	(s)	(t)	(u)	(v)	(w)	(x)	(y)	(z)
(l)	1.00														
(m)	0.78	1.00													
(n)	0.66	0.51	1.00												
(o)	0.62	0.53	0.58	1.00											
(p)	0.39	0.67	0.38	0.70	1.00										
(q)	0.05	0.03	0.22	-0.02	-0.03	1.00									
(r)	0.30	0.25	0.30	0.46	0.33	-0.08	1.00								
(s)	0.06	0.05	0.22	0.02	0.03	0.27	0.09	1.00							
(t)	0.05	-0.14	0.06	0.05	-0.23	-0.03	0.05	-0.11	1.00						
(u)	0.01	0.03	0.03	0.05	0.07	0.02	0.02	0.07	-0.06	1.00					
(v)	-0.01	0.04	-0.03	0.00	0.06	-0.01	-0.11	0.01	-0.11	0.27	1.00				
(w)	0.36	0.27	0.49	0.18	0.10	0.26	0.16	0.60	0.02	0.02	-0.10	1.00			
(x)	0.03	0.06	0.06	0.06	0.09	0.04	0.04	0.16	-0.06	0.16	0.47	0.05	1.00		
(y)	0.10	0.11	0.10	0.11	0.11	0.08	0.04	0.29	-0.05	0.08	0.25	0.05	0.41	1.00	
(z)	0.00	0.00	0.02	-0.01	0.00	0.02	0.02	0.03	0.01	-0.01	0.00	0.03	-0.01	0.03	1.00

Table 7: Response Delay Analysis (Message Level)

Variable	Response Delay (RDelay)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
MT_1	32.172***		33.874***	33.874***	33.874***	33.874***	33.874***
	(0.435)		(0.441)	(1.344)	(1.820)	(2.151)	(0.491)
$MT_1{}^2$	9.879***		7.389***	7.389***	7.389***	7.389***	7.389***
	(0.410)		(0.408)	(1.032)	(0.788)	(1.166)	(0.425)
$SkillHetMT_1$	-25.889***		-29.955***	-29.955***	-29.955***	-29.955***	-29.955***
	(0.806)		(0.827)	(2.498)	(2.955)	(3.633)	(0.791)
MT_2		14.835***	3.121***	3.121***	3.121***	3.121***	3.121***
		(0.294)	(0.317)	(0.557)	(0.500)	(0.630)	(0.317)
$MT_2{}^2$		8.714***	5.568***	5.568***	5.568***	5.568***	5.568***
		(0.211)	(0.275)	(0.462)	(0.358)	(0.469)	(0.301)
$SkillHetMT_2$		0.061	11.639***	11.639***	11.639***	11.639***	11.639***
		(0.754)	(0.801)	(1.079)	(1.185)	(1.290)	(0.756)
$log(Typing)$	13.910***	21.071***	13.817***	13.817***	13.817***	13.817***	13.817***
	(0.086)	(0.073)	(0.086)	(0.301)	(0.142)	(0.303)	(0.099)
$log(CLength)$	-2.171***	-2.748***	-2.177***	-2.177***	-2.177***	-2.177***	-2.177***
	(0.088)	(0.084)	(0.088)	(0.198)	(0.167)	(0.231)	(0.090)
$log(CSlowness)$	1.689***	1.929***	1.681***	1.681***	1.681***	1.681***	1.681***
	(0.073)	(0.076)	(0.072)	(0.102)	(0.071)	(0.091)	(0.078)
$Order(Bin=2)$	5.476***	5.928***	5.645***	5.645***	5.645***	5.645***	5.645***
	(0.162)	(0.172)	(0.162)	(0.372)	(0.280)	(0.413)	(0.162)
$Order(Bin=3)$	-7.749***	-9.113***	-7.304***	-7.304***	-7.304***	-7.304***	-7.304***
	(0.201)	(0.211)	(0.201)	(0.482)	(0.480)	(0.621)	(0.213)
Intercept	15.381***	-7.685***	10.845***	10.845***	10.845***	10.845***	10.845***
	(0.372)	(0.420)	(0.451)	(1.229)	(0.603)	(1.212)	(0.425)
Session f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Standard errors	Robust	Robust	Robust	Clustered (agent)	Clustered (date)	Clustered (agent and date)	Bootstrapped
No. of obs.	600,376	600,376	600,376	600,376	600,376	600,376	600,376
R-squared	0.415	0.384	0.416	0.416	0.416	0.416	0.416

1. $*p < 0.1, **p < 0.05, ***p < 0.01$
2. MT values are centered.
3. Columns 1 and 2 use the two different multitasking / associated variables individually. The remaining models use all the multitasking variables together.
4. Robust standard errors are shown in parentheses. Standard errors are clustered at the agent level in column 4, at the date level in column 5 and at both agent and date levels in column 6. Bootstrapped standard errors are used in column 7 (100 sample bootstrapping).
5. For two-way clustering, we use the procedure at: http://acct.wharton.upenn.edu/~dtayl/code.htm

Table 8: Problem Resolution Analysis (Session Level)

Variable	Resolution (Survey)				
	(1)	(2)	(3)	(4)	(5)
$AvgMT_1$	-0.747***		-0.724***	-0.724***	-0.724***
	(0.120)		(0.148)	(0.147)	(0.179)
$AvgSkillHetMT_1$	0.341		0.285	0.285	0.285
	(0.277)		(0.373)	(0.367)	(0.409)
$sdMT_1$	0.098		0.138	0.138	0.138
	(0.105)		(0.110)	(0.109)	(0.120)
$AvgMT_2$		-0.244***	-0.053	-0.053	-0.053
		(0.064)	(0.079)	(0.080)	(0.054)
$AvgSkillHetMT_2$		0.176	0.036	0.036	0.036
		(0.139)	(0.189)	(0.199)	(0.200)
$sdMT_2$		-0.100	-0.120	-0.120	-0.120
		(0.089)	(0.091)	(0.088)	(0.074)
$log(QueueWait)$	-0.028*	-0.029*	-0.028*	-0.028*	-0.028
	(0.015)	(0.016)	(0.016)	(0.016)	(0.018)
$log(Duration)$	-0.992***	-0.947***	-0.988***	-0.988***	-0.988**
	(0.320)	(0.320)	(0.320)	(0.330)	(0.396)
$log(AExperience)$	-0.052	-0.043	-0.050	-0.050	-0.050
	(0.034)	(0.034)	(0.034)	(0.033)	(0.033)
$log(NegEmotion)$	-0.317***	-0.318***	-0.318***	-0.318***	-0.318***
	(0.082)	(0.082)	(0.082)	(0.082)	(0.09)
$log(DescLength)$	-0.394	-0.425	-0.398	-0.398	-0.398
	(0.275)	(0.274)	(0.275)	(0.280)	(0.322)
$log(TotalTyping)$	0.468***	0.383***	0.471***	0.471***	0.471***
	(0.041)	(0.036)	(0.042)	(0.048)	(0.042)
$log(AvgCLength)$	-0.364***	-0.350***	-0.365***	-0.365***	-0.365***
	(0.048)	(0.048)	(0.048)	(0.048)	(0.039)
$log(AvgCSlowness)$	0.105**	0.075	0.108**	0.108**	0.108***
	(0.049)	(0.049)	(0.049)	(0.055)	(0.035)
$log(SurveyGap)$	0.159***	0.163***	0.160***	0.160***	0.160***
	(0.048)	(0.048)	(0.048)	(0.051)	(0.052)
λ	-3.730*	-3.867*	-3.751*	-3.751*	-3.751
	(2.053)	(2.048)	(2.052)	(2.101)	(2.457)
Agent f.e.	Yes	Yes	Yes	Yes	Yes
Skill key f.e.	Yes	Yes	Yes	Yes	Yes
Product key f.e.	Yes	Yes	Yes	Yes	Yes
Contact date f.e.	Yes	Yes	Yes	Yes	Yes
Contact time f.e.	Yes	Yes	Yes	Yes	Yes
Standard errors	Robust	Robust	Robust	Clustered (agent)	Clustered (date)
No. of obs.	13,507	13,507	13,507	13,507	13,507

1. *p < 0.1, **p < 0.05, ***p < 0.01
2. *Columns 1 and 2 use the two different multitasking / associated variables individually. The remaining models use all the multitasking variables together.*
3. *Robust standard errors are shown in parentheses. Standard errors are clustered at the agent level in column 4, at the date level in column 5.*

Table 9: Effects of Multitasking on Expectation Fulfillment via Average Response Delay and Resolution

Panel A: Regression Analyses

Variable	Avg. Response Delay ($AvgRDelay$)			Resolution ($Resolution$)			Expectation Fulfillment ($Satisfaction_1$)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$AvgMT_1$	63.807***		49.788***	-0.150***		-0.053*	-0.064		0.007
	(1.390)		(1.748)	(0.024)		(0.031)	(0.053)		(0.066)
$AvgSkillHetMT_1$	-30.365***		-51.799***	0.064		0.072	-0.060		-0.119
	(3.169)		(4.127)	(0.054)		(0.073)	(0.112)		(0.152)
$sdMT_1$	-13.149***		-23.093***	0.020		0.111***	0.054		0.155***
	(1.187)		(1.176)	(0.020)		(0.021)	(0.042)		(0.044)
$AvgMT_2$		16.031***	-1.408*		-0.047***	-0.018		-0.030	-0.051
		(0.748)	(0.754)		(0.012)	(0.015)		(0.026)	(0.032)
$AvgSkillHetMT_2$		-4.475***	17.769***		0.034	-0.008		0.013	0.041
		(1.725)	(2.059)		(0.027)	(0.036)		(0.056)	(0.076)
$sdMT_2$		2.357**	1.953**		-0.020	-0.002		-0.081**	-0.078**
		(1.093)	(0.988)		(0.017)	(0.017)		(0.035)	(0.036)
$log(QueueWait)$	-0.025	-0.035	-0.130	-0.005*	-0.006*	-0.005*	-0.020***	-0.021***	-0.021***
	(0.151)	(0.168)	(0.149)	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)
$log(Duration)$	10.782***	5.795***	10.817***	-0.208***	-0.196***	-0.167***	-0.091	-0.088	-0.035
	(0.458)	(0.492)	(0.365)	(0.059)	(0.059)	(0.059)	(0.124)	(0.124)	(0.124)
$log(AExperience)$	-1.993***	-2.321***	-1.773***	-0.010	-0.008	-0.009	-0.021	-0.020	-0.019
	(0.375)	(0.409)	(0.362)	(0.006)	(0.006)	(0.006)	(0.013)	(0.013)	(0.013)
$log(NegEmotion)$	-0.307	-0.725	-0.228	-0.067***	-0.066***	-0.070***	-0.003	-0.004	-0.007
	(0.528)	(0.577)	(0.510)	(0.015)	(0.015)	(0.015)	(0.032)	(0.032)	(0.032)
$log(DescLength)$	2.357***	3.383***	2.403***	-0.088*	-0.091*	-0.106**	-0.079	-0.088	-0.105
	(0.468)	(0.511)	(0.451)	(0.051)	(0.051)	(0.051)	(0.106)	(0.106)	(0.106)
$log(TotalTyping)$	3.134***	3.687***	18.797***	0.094***	0.077***	-0.044***	0.106***	0.108***	0.021
	(0.448)	(0.438)	(0.596)	(0.008)	(0.007)	(0.011)	(0.016)	(0.014)	(0.023)
$log(AvgCLength)$	1.959***	0.725	1.420***	-0.072***	-0.069***	-0.065***	-0.096***	-0.094***	-0.089***
	(0.524)	(0.571)	(0.505)	(0.009)	(0.009)	(0.009)	(0.019)	(0.019)	(0.019)
$log(AvgCSlowness)$	2.335***	4.872***	1.814***	0.023**	0.016*	0.004	-0.001	-0.001	-0.025
	(0.553)	(0.600)	(0.526)	(0.009)	(0.009)	(0.009)	(0.020)	(0.019)	(0.019)
$log(SurveyGap)$				0.030***	0.030***	0.031***	0.007	0.007	0.008
				(0.009)	(0.009)	(0.009)	(0.018)	(0.018)	(0.018)
λ				-0.810**	-0.817**	-0.865**	-0.718	-0.776	-0.840
				(0.378)	(0.378)	(0.380)	(0.794)	(0.793)	(0.795)
$AvgRDelay$							-0.002***	-0.002***	-0.002***
							(0.000)	(0.000)	(0.000)
$Resolution$							2.003***	2.003***	2.014***
							(0.018)	(0.018)	(0.018)
Agent f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Skill key f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Product key f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Contact date f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Contact time f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	14,021	14,021	14,021	14,021	14,021	14,021	14,021	14,021	14,021

Panel B: Path Analysis

Path	$AvgMT_1$	$AvgMT_2$	Both MTs together	
			$AvgMT_1$	$AvgMT_2$
Indirect effect via *Response Delay*	-0.097***	-0.027***	-0.083***	0.002
	(0.019)	(0.005)	(0.016)	(0.001)
Indirect effect via *Resolution*	-0.300***	-0.095***	-0.107*	-0.037
	(0.047)	(0.024)	(0.062)	(0.031)
Total indirect eff.	-0.396***	-0.122***	-0.190***	-0.035
	(0.051)	(0.025)	(0.064)	(0.031)
Direct effect	-0.064	-0.030	0.007	-0.051
	(0.053)	(0.026)	(0.066)	(0.032)
Total effect	-0.460***	-0.152***	-0.183**	-0.086*
	(0.068)	(0.035)	(0.089)	(0.044)

Panel A: Regression Analyses

Variable	Avg. Response Delay (AvgRDelay)			Resolution (Resolution)			Promotion (Satisfaction$_2$)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$AvgMT_1$	63.679***		70.291***	-0.151***		-0.147***	-0.077*		-0.072
	(1.390)		(1.708)	(0.024)		(0.029)	(0.040)		(0.048)
$AvgSkillHetMT_1$	-30.706***		-52.979***	0.069		0.056	-0.100		0.069
	(3.171)		(4.267)	(0.054)		(0.072)	(0.085)		(0.115)
$sdMT_1$	-12.982***		-15.439***	0.022		0.029	0.063*		0.078**
	(1.186)		(1.249)	(0.020)		(0.021)	(0.032)		(0.034)
$AvgMT_2$		16.040***	-1.969**		-0.048***	-0.009		-0.015	-0.013
		(0.784)	(0.890)		(0.012)	(0.015)		(0.019)	(0.024)
$AvgSkillHetMT_2$		-4.480***	16.866***		0.035	-0.008		0.047	0.021
		(1.724)	(2.129)		(0.027)	(0.036)		(0.042)	(0.057)
$sdMT_2$		2.403**	4.465***		-0.019	-0.023		-0.048**	-0.062**
		(1.092)	(1.021)		(0.017)	(0.017)		(0.027)	(0.027)
$log(QueueWait)$	-0.027	-0.026	-0.096	-0.005*	-0.006*	-0.006*	-0.012**	-0.012***	-0.012***
	(0.151)	(0.168)	(0.153)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
$log(Duration)$	10.789***	5.834***	10.459***	-0.204***	-0.192***	-0.203***	-0.132	-0.130	-0.129
	(0.457)	(0.492)	(0.461)	(0.059)	(0.059)	(0.059)	(0.094)	(0.094)	(0.094)
$log(AExperience)$	-1.984***	-2.323***	-1.864***	-0.010	-0.009	-0.010	-0.025**	-0.025**	-0.025**
	(0.374)	(0.409)	(0.374)	(0.006)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)
$log(NegEmotion)$	-0.359	-0.736	-0.393	-0.066***	-0.065***	-0.066***	-0.064***	-0.064***	-0.064***
	(0.528)	(0.576)	(0.527)	(0.015)	(0.015)	(0.015)	(0.024)	(0.024)	(0.024)
$log(DescLength)$	2.407***	3.437***	2.485***	-0.085*	-0.088*	-0.085*	-0.054	-0.058	-0.055
	(0.468)	(0.510)	(0.467)	(0.051)	(0.051)	(0.051)	(0.080)	(0.080)	(0.080)
$log(TotalTyping)$	3.190***	3.598***	3.071***	0.094***	0.077***	-0.094***	0.024**	0.030***	0.026**
	(0.448)	(0.438)	(0.451)	(0.008)	(0.007)	(0.008)	(0.012)	(0.011)	(0.012)
$log(AvgCLength)$	1.892***	0.659	1.420***	-0.072***	-0.069***	-0.072***	-0.077***	-0.076***	-0.077***
	(0.524)	(0.571)	(0.505)	(0.009)	(0.009)	(0.009)	(0.014)	(0.014)	(0.014)
$log(AvgCSlowness)$	2.335***	4.841***	1.880***	0.022**	0.016*	0.022**	-0.016	-0.017	-0.017
	(0.553)	(0.600)	(0.522)	(0.009)	(0.009)	(0.009)	(0.015)	(0.015)	(0.015)
$log(SurveyGap)$				0.031***	0.031***	0.031***	0.004	0.005	0.005
				(0.009)	(0.009)	(0.009)	(0.014)	(0.014)	(0.014)
λ				-0.785**	-0.793**	-0.841**	-0.730	-0.760	-0.738
				(0.378)	(0.378)	(0.339)	(0.601)	(0.601)	(0.601)
$AvgRDelay$							-0.002***	-0.001***	-0.001***
							(0.000)	(0.000)	(0.000)
$Resolution$							0.998***	0.998***	0.998***
							(0.013)	(0.013)	(0.013)
Agent f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Skill key f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Product key f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Contact date f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Contact time f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	14,069	14,069	14,069	14,069	14,069	14,069	14,069	14,069	14,069

Panel B: Path Analysis

Path	$AvgMT_1$	$AvgMT_2$	Both MTs together $AvgMT_1$	$AvgMT_2$
Indirect effect via *Response Delay*	-0.024*	-0.008**	-0.026*	0.001
	(0.014)	(0.003)	(0.016)	(0.001)
Indirect effect via *Resolution*	-0.151***	-0.048***	-0.147***	-0.009
	(0.024)	(0.012)	(0.029)	(0.015)
Total indirect eff.	-0.175***	-0.056***	-0.173***	-0.008
	(0.028)	(0.013)	(0.033)	(0.015)
Direct effect	-0.077*	-0.015	-0.072	-0.013
	(0.040)	(0.019)	(0.050)	(0.024)
Total effect	-0.252***	-0.071***	-0.245***	-0.021
	(0.044)	(0.023)	(0.054)	(0.028)

Table 11: Propensity Score Matching Results for Avg. Resp. Delay at the Session Level

Threshold to be Considered as Treated	Multitasking variable	ATET	95% Conf. Interval (Low)	95% Conf. Interval (High)
1.3	$AvgMT_1$	21.871*** (0.330)	21.223	22.518
1.4	$AvgMT_1$	24.062*** (0.350)	23.375	24.749
1.5	$AvgMT_1$	27.985*** (0.427)	27.149	28.821
1.6	$AvgMT_1$	29.788*** (0.451)	28.904	30.672
1.7	$AvgMT_1$	33.222*** (0.582)	32.080	34.363
1.8	$AvgMT_1$	36.486*** (0.613)	35.192	38.116
1.3	$AvgMT_2$	14.774*** (0.795)	13.215	16.333
1.4	$AvgMT_2$	13.214*** (0.644)	11.952	14.475
1.5	$AvgMT_2$	11.468*** (0.671)	10.152	12.783
1.6	$AvgMT_2$	11.099*** (0.662)	9.801	12.397
1.7	$AvgMT_2$	9.996*** (0.518)	8.980	11.011
1.8	$AvgMT_2$	8.564*** (0.474)	7.635	9.493

Table 12: Propensity Score Matching Results for Problem Resolution at the Session Level

Threshold to be Considered as Treated	Multitasking variable	ATET	95% Conf. Interval (Low)	95% Conf. Interval (High)
1.3	$AvgMT_1$	-0.013 (0.009)	-0.032	0.005
1.4	$AvgMT_1$	-0.016 (0.010)	-0.036	0.004
1.5	$AvgMT_1$	-0.044*** (0.013)	-0.069	-0.020
1.6	$AvgMT_1$	-0.058*** (0.009)	-0.077	-0.039
1.7	$AvgMT_1$	-0.085*** (0.003)	-0.091	-0.078
1.8	$AvgMT_1$	-0.092*** (0.013)	-0.117	-0.067
1.3	$AvgMT_2$	0.016 (0.016)	-0.015	0.047
1.4	$AvgMT_2$	0.007 (0.012)	-0.016	0.030
1.5	$AvgMT_2$	-0.014 (0.018)	-0.049	0.021
1.6	$AvgMT_2$	-0.002 (0.014)	-0.030	0.026
1.7	$AvgMT_2$	0.002 (0.013)	-0.024	0.028
1.8	$AvgMT_2$	0.004 (0.011)	-0.018	0.026

Table 13: Paired t-test Results after Matching at the Message Level

Variable	Mean	N	Std. err.	95% Conf. Int. (Low)	95% Conf. Int. (High)	t	Sig.
MT	69.026	34,382	0.283	68.471	69.582		
No MT	52.876	34,382	0.274	52.339	53.412		
Difference	16.150	34,382	0.371	15.423	16.878	43.503	0.000

Table 14: Two Segment Finite Mixture Model

	Expectation Fulfillment ($Satisfaction_1$)			
	(1)		(2)	
	1-Class Model		2-Class Model	
AIC (LL)	43,524.57		42,554.21	
BIC (LL)	44,000.16		43,796.87	
CAIC (LL)	44,063.16		43,961.87	
R-squared	0.076		0.233	
	Segment	Level	Segments	
	1		1	2
Predictors				
$AvgMT_1$	-0.268***		-0.439***	-0.215***
	(0.039)		(0.082)	(0.071)
$AvgSkillHetMT_1$	0.029		0.107	0.060
	(0.087)		(0.184)	(0.163)
$sdMT_1$	0.049		0.037	0.054
	(0.033)		(0.073)	(0.057)
$log(QueueWait)$	-0.013***		-0.017*	-0.005
	(0.004)		(0.008)	(0.007)
$log(Duration)$	-0.129***		-0.151***	-0.141***
	(0.012)		(0.036)	(0.024)
$log(AExperience)$	0.025***		0.030**	0.035***
	(0.006)		(0.012)	(0.011)
$log(NegEmotion)$	-0.024*		0.066	-0.088***
	(0.015)		(0.040)	(0.027)
$log(DescLength)$	0.027**		0.153***	-0.061**
	(0.013)		(0.033)	(0.029)
$log(TotalTyping)$	0.230***		0.349***	0.181***
	(0.012)		(0.035)	(0.023)
$log(AvgCLength)$	-0.153***		-0.280***	-0.065**
	(0.015)		(0.035)	(0.030)
$log(AvgCSlowness)$	0.015		0.013	0.039
	(0.015)		(0.035)	(0.029)
$log(SurveyGap)$	0.042***		0.034	0.059**
	(0.014)		(0.029)	(0.028)
Skill key f.e.	Yes		Yes	Yes
Contact date f.e.	Yes		Yes	Yes
Model For Segments				
Gender	N/A		0.132***	-0.132***
			(0.039)	(0.039)
Skill domain	N/A	Getting started	-0.274**	0.274**
			(0.109)	(0.109)
	N/A	Shop/buy	0.851***	-0.851***
			(0.163)	(0.163)
	N/A	Working on file	-0.576***	0.576***
			(0.136)	(0.136)
Product key	N/A	19 levels	Yes	Yes
Contact time	N/A	18 levels	Yes	Yes
Relative segment size	1.00		0.51	0.49