12-2017

Call Center Experience Optimization: A Case for a Virtual Predictive Queue

William Pugh
University of the Incarnate Word, wpugh@uiwtx.edu

CALL CENTER EXPERIENCE OPTIMIZATION: A CASE FOR A VIRTUAL PREDICTIVE QUEUE

by

WILLIAM K. PUGH

A DISSERTATION

Presented to the Faculty of the University of the Incarnate Word
in partial fulfillment of the requirements
for the degree of

DOCTOR OF BUSINESS ADMINISTRATION

UNIVERSITY OF THE INCARNATE WORD

August 2017

ACKNOWLEDGEMENTS

I wish to express my extreme gratitude to several individuals who have enabled my success as a student and the successful completion of this dissertation. First, I would like to thank my dissertation committee including Dr. Alicia Rodriguez de Rubio (University of the Incarnate Word, San Antonio, Texas) and Dr. Joleen Beltrami (University of the Incarnate Word, San Antonio, Texas). I would also like to give special acknowledgement to Dr. Vess Johnson (University of the Incarnate Word, San Antonio, Texas) for his dedication and support.

I would also like to thank Professor Avishai Mandelbaum (Technion – Israel Institute of Technology, Haifa, Israel) for sharing his academic research on queueing theory and his correspondence in providing a comprehensive bibliography reference list. I am also grateful to Dr. Susan Torres-Harding (Roosevelt University, Chicago, Illinois) for her research on social justice and providing valuable information relative to my study. In addition, I would like to thank my fellow doctoral candidates Tiffany Wetz and Tom Dotter for their encouragement throughout our academic program, staying focused on the objective, and remaining strong throughout many surprises and disappointments.

Finally, I would like to thank my family including my children (Samantha, Julia, and Garner) and my wife René for remaining patient and understanding while I pursue this great academic endeavor.

William K. Pugh

CALL CENTER EXPERIENCE OPTIMIZATION: A CASE
FOR A VIRTUAL PREDICTIVE QUEUE

William K. Pugh, Doctor of Business Administration

University of the Incarnate Word, 2017

The evolution of the call center into contact centers and the growth of their use in providing customer-facing service by many companies has brought considerable capabilities in maintaining customer relationships but it also has brought challenges in providing quality service when call volumes are high. Limited in their ability to provide service at all times to all customers, companies are forced to balance the costs associated with hiring more customer service representatives and the quality of service provided by a fewer number. A primary challenge when there are not enough customer service representatives to engage the volume of callers in a timely manner is the significant wait times that can be experienced by many customers. Normally, callers are handled in accordance with a first-come, first-served policy with exceptions being skill-based routing to those customer service representatives with specialized skills.

Queuing theory applies discrete mathematical principles in the study of queues, or waiting lines. Queuing theory formulas are used to determine operating characteristics of a queue such as the probability no customers are in the queue, the average number of customers waiting in the queue and the average time a customer will spend in the entire service system. These principles are used today to predict arrival behavior of customers into a service queue. When these assumptions are violated due to an unanticipated larger than normal call volume or a

reduced number of customer service representatives based on inaccurate forecasts, the call center service queue wait times become increasingly large and can cause an overall negative customer experience. A proposed call center infrastructure framework called a Virtual Predictive Queue (VPQ) can allow some customers to benefit from a shorter call queue wait time. This proposed system can be implemented within a call center's Automatic Call Distribution device associated with computer telephony integration. A key factor in the proposed VPQ integration is that the servicer can decide who can enter the VPQ. Another important feature of the proposed VPQ infrastructure is that it does not violate the common first-in, first-out policy. The advanced reservation feature of the VPQ can be accounted for when providing customers in the normal service queue with an expected wait time. Deciding how many advanced reservations that should be created within the VPQ can be based on predictive analytics, past performance or can be invoked real-time when arrivals begin to increase past expected volumes.

The fundamental problem a VPQ can address is the extremely long call queue wait time experienced by some customers. Long wait times can negatively affect a person's perceived overall service experience by exceeding an individual's patience threshold where they can choose to abandon a queue or continue to hold until served. The impact of exceeding customer patience thresholds can form the basis for a strong business case for implementing a VPQ. Understanding the impact of customer patience on service queue abandonment can lead to better customer service and long-term satisfaction.

This study found a significant relationship between customer patience and intent to abandon a service queue. In addition, perceived justice within a service queue had an impact on whether some customers choose to abandon a queue or not. Investigation into both apparent (line

standing) and phantom (call queues) discovered that both perceived justice and exceeding a person's patience threshold contributed to their intent to abandon the queue.

This study also found that invoking a VPQ with advanced reservations through a simulation experiment contributed to a significant reduction in wait time between customers occupying a normal service queue and those in the VPQ. A discrete event simulation using a spreadsheet application found significant time savings in 10 consecutive simulation executions. The intent was to demonstrate that invoking advanced reservations associated with a VPQ can reduce wait time for a select customer population. The findings found within this investigation can contribute to a better understanding of customer perceptions of waiting in a queue and their intent to abandon. Additionally, the demonstrated performance of a simulated VPQ offers a positive outlook on its effectiveness for reducing wait times for a select customer group.

TABLE OF CONTENTS

Table of Contents—Continued

Table of Contents—Continued

Table of Contents—Continued

LIST OF TABLES

LIST OF FIGURES

Research Background

Context of the Study

Contact center growth is steadily increasing and has become a valuable business tool for building strong customer relationships (Deloitte, 2015). A contact center is a natural evolution of what was once termed a call center where businesses traditionally engaged customers during inbound call operations (Holman, Batt, & Holtgrewe, 2007). In addition to providing service over the phone channel, the modern contact center now integrates other forms of direct customer contact such as mail, e-mail, social media, and online chat capability in one location (Ali, 2010; Koole & Pot, 2006). However, contact centers are not without challenges. The costs associated with operating and maintaining contact centers are significant with the majority of the budget being spent on human resources and staffing (Akhtar & Latif, 2010; Aksin, Armony, & Mehrotra, 2007; Weinberg, Brown, & Stroud, 2007)

Additional considerations are balancing staffing levels while providing quality customer service, particularly during high-volume traffic areas where customers may find themselves waiting for extended periods to speak to a customer service representative (CSR). Customer frustration arises when wait times exceed a person's expectation for service (Maister, 2005). Excessive wait times develop because of several factors, such as understaffed call centers, unanticipated call volumes and excessive average handle time during customer service (Gans, Koole, & Mandelbaum, 2003). Regardless of the reason, callers who wait in a queue can perceive their waits differently and this phenomenon has been studied from a social justice perspective (Larson, 1987). Social justice implications can have a profound effect on customer service experience when an injustice occurs.

Rawls (1971) offers a definition of social justice in which he equates social justice as that of *justice as fairness*. He also proposes a thought experiment where individuals who are deciding on what type of society to create for themselves are first placed in an *original position* where no knowledge exists about individual social status, strengths, or limitations (Rawls, 1971). Under this *veil of ignorance*, Rawls (1971) asserts that all individuals would choose a society where rules would be impartial in their application and resources distributed equally. Any failure in the proposed framework would create an injustice to one or more individuals. The impact of waiting in a queue, and the perceived injustice that can arise, comes when someone who entered the queue after a person who is already waiting, but is served before them.

To impart a level of fairness to all customers, the typical policy for servicing customers waiting in the queue is first-come, first-served (FCFS) and call center infrastructure follows this principle by implementing a first-in, first-out (FIFO) policy for callers (Gans et al., 2003; Koole & Mandelbaum, 2002). Call centers must consider trade-offs between efficiency and agent costs when making staffing considerations. More agents usually will increase service quality, but costs can accumulate when some agents are idle or even underutilized (Ali, 2010; Fukunaga et al., 2002).

The study of waiting lines, or queues, is called *queuing theory* and is considered a branch of operations research and continued research is used to make business decisions involving allocating resources when providing service. The origins of queuing theory begin when A.K. Erlang proposed waiting models for the Copenhagen telephone exchange at the beginning of the twentieth century (Aksin et al., 2007). Erlang’s influential work resulted in several publications that describe call arrivals as a *Poisson* process. A Poisson process is a probability distribution that describes the given number of events occurring over a fixed interval of time provided the

average arrival rate is known. In addition, a Poisson process is considered *memoryless* in that each arrival is independent of the last arrival (Gans, Liu, Mandelbaum, Shen, & Han Ye, 2010; Koole & Pot, 2006). Erlang models are written in *Kendall* notation in the form A/S/c where A denotes time between arrivals in a queue, S is the service time distribution and c is the number of servers (Sankaranarayanan, 2011). The classical queuing model is the M/M/1, or Erlang-C model (Borst, Mandelbaum, & Reiman, 2004; Brown et al., 2005; Gans et al., 2010). This form denotes both arrival and service times are Markovian (M) with a single server (Brown et al., 2005).

Call center telephony infrastructure has advanced over the years and now incorporates many technology changes to facilitate customer contact and service resolution. For instance, Interactive Voice Response (IVR) allows customers to complete certain transactions by pressing numbers on the keypad and sending Dual Tone, Multi Frequency tones or by spoken word (Mathew & Nambiar, 2013). This capability can reduce the number of customers who must wait in a queue to speak to a CSR. Customers who desire to speak to a CSR are routed through a device known as an automatic call distributor. The automatic call distributor then routes the calls based on predetermined criteria. For instance, calls may be routed by desired service type or to CSRs with specific skill sets (Aksin et al., 2007; Koole & Pot, 2006; Wallace & Whitt, 2005). The automatic call distributor maintains a record of each CSR's skill set and routes calls to the appropriate agent when they are logged on and idle. Agent utilization is normally measured over short intervals (half-hours) during each day and is calculated as the average number of CSRs that were active during that period (Aksin et al., 2007; Garnett, Mandelbaum, & Reiman, 2002; Shen, 2010). CSR utilization is one key metric in call center operations, and combined with the average handle time information, call center managers use this information to anticipate future staffing levels and gauge call center performance (Aksin et al., 2007; Bouzada, 2009). When agents

become saturated with calls, customers are placed on hold and the queue can build exponentially if the average handle time is greater than the call arrival rate. Although customer arrivals can be averaged over a specific period, they follow a Poisson probability distribution whose function is indicated by *Equation 1*

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2 \ldots \tag{1}$$

where x is the number of arrivals and λ is the *mean* number of arrivals in the time period with $e = 2.71828$. The Poisson cumulative distribution function is used to calculate the probability of the number of arrivals being less than or equal to x and is indicated by *Equation 2*

$$F(x;\lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}\,. \tag{2}$$

The service time represented by the M/M/1 queue is exponentially distributed and the probability density function (PDF) is indicated by *Equation 3*

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0\,. \end{cases} \tag{3}$$

Integrating *Equation* 3 gives the cumulative distribution function which is shown in *Equation 4*

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0\,. \end{cases} \tag{4}$$

This gives the cumulative probability of having a service duration of x given arrival rate λ.

Call center queues typically have three major designs for queuing incoming calls. The first is shown in Figure 1. In this model, a customer has no idea how long the wait time will be but chooses to remain in the queue for service and W_s is where waiting in the queue ends and waiting for service begins.

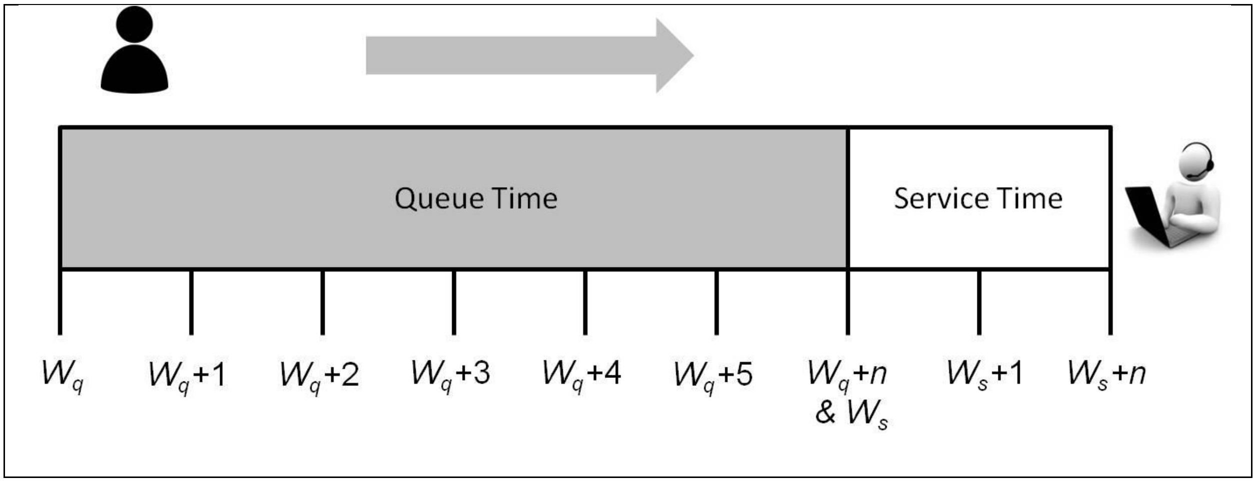


Figure 1. Normal call queue with unknown wait times (Pugh, 2016).

The second model, shown in Figure 2, provides the customer with a recorded message on how long the expected wait will be.

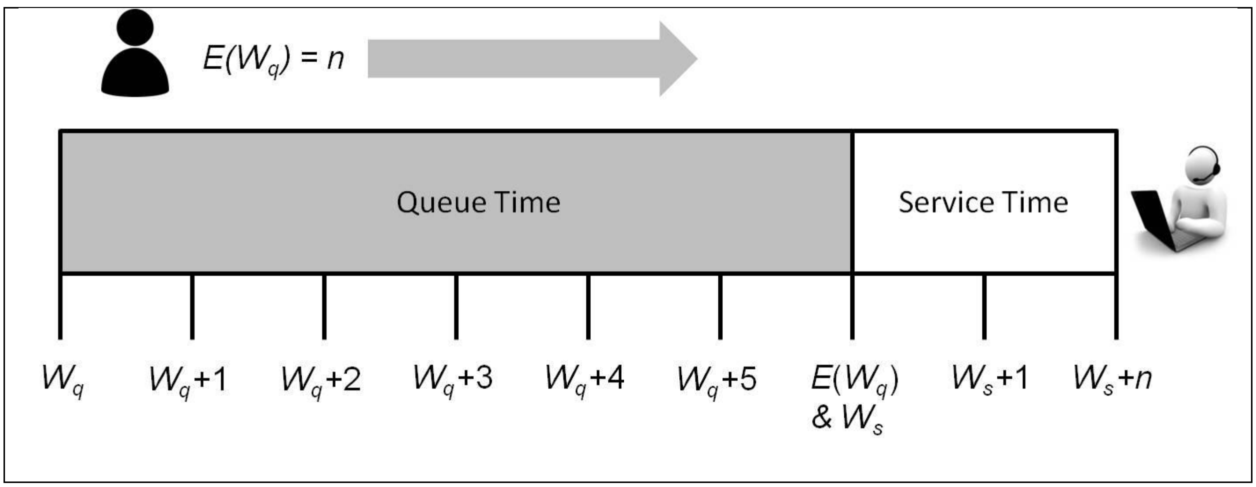


Figure 2. Normal call queue with announced wait times (Pugh, 2016).

Realistically, the announced expected wait time $E(W_q)$ for service will be close to the actual wait time for the customer and where service begins. The customer chooses whether to remain in the queue or abandon the call. The third model, shown in Figure 3, is a virtual queue where the customer is given the option to wait with an expected wait time or to be called back when a CSR is free.

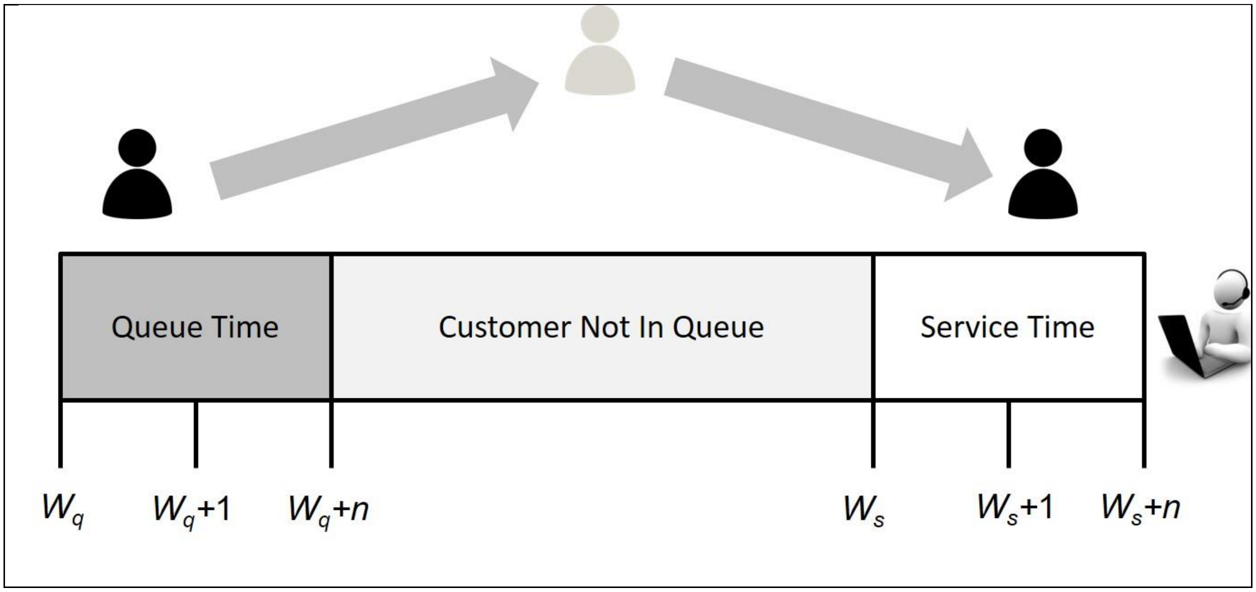


Figure 3. Call queue with callback capability (Pugh, 2016).

Statement of the Problem

Customers subjected to long wait times can become frustrated and their perception of the service quality may be negatively affected by that experience (Taylor, 1994). When customers' expectations are not met, the psychological impact can have profound effects on their perception and feedback on the quality of service (Larson, 1987; Maister, 2005). In his book *The Psychology of Waiting Lines*, (Maister, 2005) presents a hypothetical satisfaction formula in the form $S = P – E$, where S is satisfaction, P is perception and E is expectation. Realistically, if a person's perception were greater than their expectation, a positive satisfaction score would result. Conversely, if the expectation were greater than the perception, a negative score would result. Human impatience is a natural behavior exhibited by individuals when forced to wait in a queue for service (Gans et al., 2003). Studies on impatience have produced interesting results on the phenomenon of waiting. Gans et al. (2003) also explore impatience and impart a quantitative understanding through the application of an impatience function of how individuals respond to waiting for service.

What mechanisms exist to reduce wait times for callers in a queue who may warrant quicker service? A common adage states: “When everyone is special, nobody is special” that intuitively suggests not everyone can be afforded priority service in all situations. However, what can be done to alleviate the excessive wait times sometimes associated with call center operations? Logically, increasing the number of CSRs that are available would go far in reducing wait times. However, this would also increase costs for a business and ultimately affect the bottom line, even for a profitable endeavor. Another option is to block all callers who must wait for service since and they would effectively get a busy signal when calling (Aksin et al., 2007). This operation would undoubtedly receive poor ratings from customers and lead to a negative experience as well. If businesses could alleviate the wait times for a select group of customers, faster service would mean higher perceived service quality and a better customer experience for those customers. With these limitations in mind, a *Virtual Predictive Queue* (VPQ) can be utilized to service a select group from a population of customers. A VPQ can be utilized in several diverse ways to bring value to individuals, organizations, and businesses. The following applications and models illustrate potential uses of the VPQ. The Person to Business (P2B) Application, Internal to Business (INT2B) Model in depicted in Figure 4.

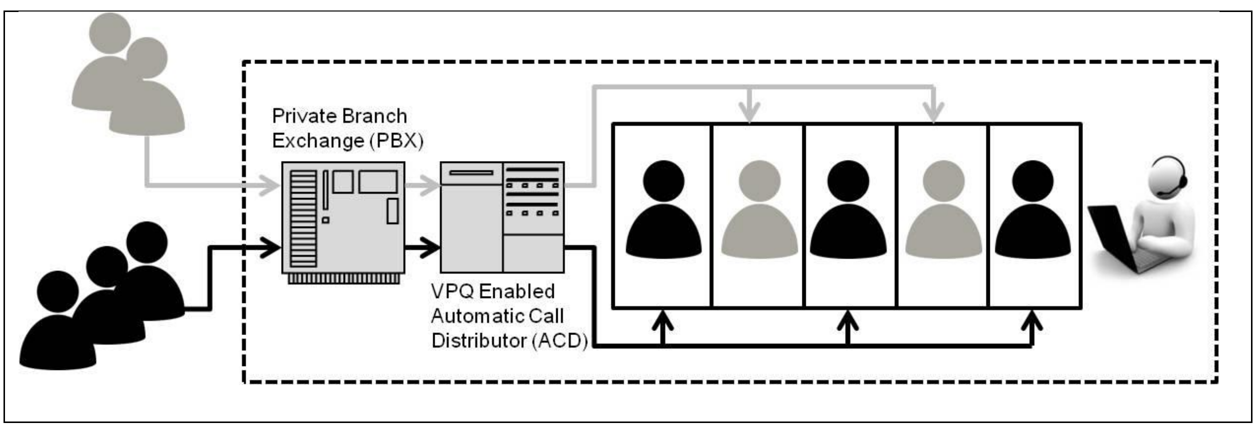


Figure 4. P2B application, INT2B model for VPQ (Pugh, 2016).

In this configuration, the VPQ is enabled within the automatic call distributor inside an organization's infrastructure. This is meant to allow a deliberate attempt to provide faster service times to a select population based on server-defined requirements. The customers to utilize the VPQ can be determined within the telephone infrastructure by resolving phone number or other information. This could also be achieved through a VOIP gateway for those calls originating from a computer or an application on a mobile device. Making VOIP calls through the mobile device, geographic location could be resolved and customers in certain areas may be allowed to utilize the VPQ for faster service in cases such as natural disasters. An example could be when customers want to submit insurance claims immediately due to a natural disaster or in response to a targeted marketing campaign. This configuration is compatible with all queue models and does not violate a traditional FIFO policy in that the advanced reservation made by the VPQ is considered when announcing estimated wait times to customers. By not preempting customers already in a queue, this will allow for faster service for those customers with a prioritized need and not impact customers waiting in the service queue. The Person to Business (P2B) Application, External to Business (EXT2B) Model is depicted in Figure 5. In this configuration, the VPQ could become enabled as a service provided to customers who are willing to pay a fee for use. The service could be used to target businesses and organizations with extremely long call wait times. In this application, the VPQ must be able to navigate the IVR system with the appropriate words or tones to allow direct entry into the queue. This service could also be delivered through a mobile application with selections made for individual businesses or organizations to call for service. The software would also be able to provide up-to-the-minute call wait times and predicted time saved if using the VPQ service based on real-time call statistics.

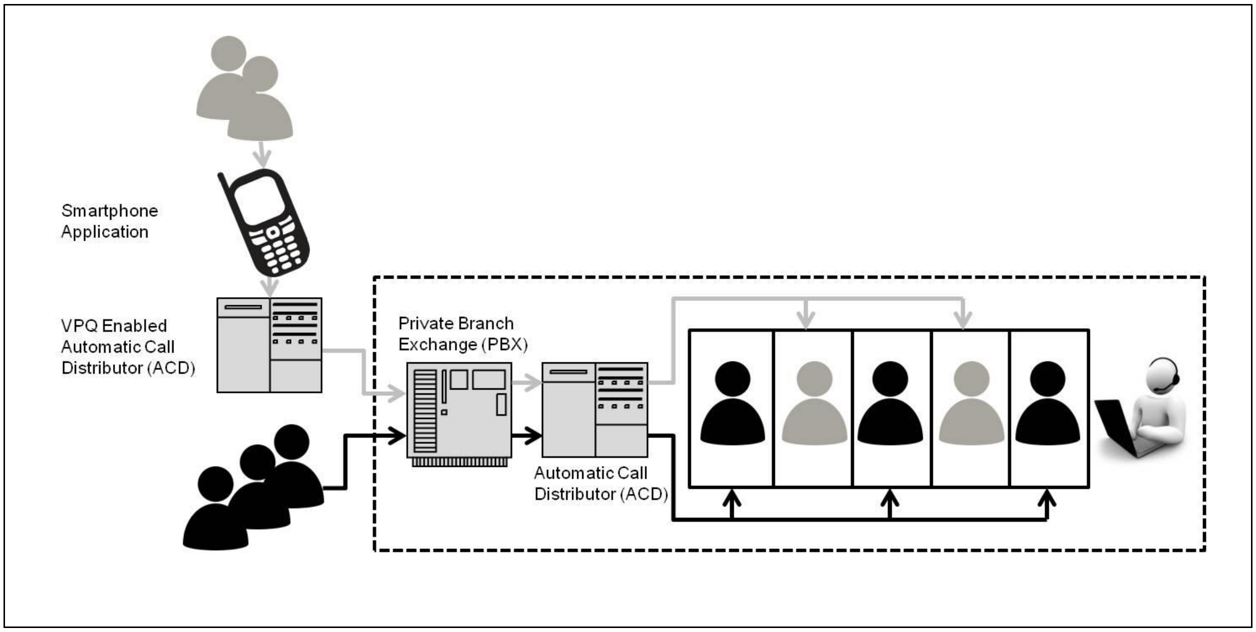


Figure 5. P2B application, EXT2B model for VPQ (Pugh, 2016).

The Business-to-Business (B2B) Application, Internal to Business (INT2B) Model is shown in Figure 6.

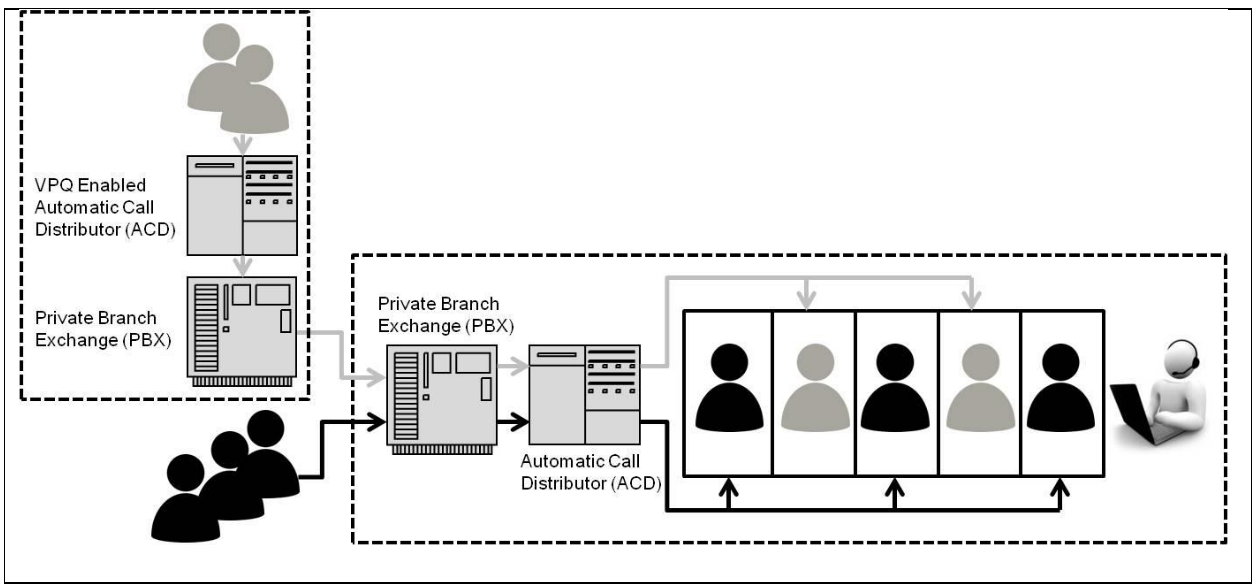


Figure 6. B2B application, INT2B model for VPQ (Pugh, 2016).

In this model, the VPQ could be used by businesses to target other organizations or businesses that have long call wait times. This could be used by back office personnel conducting regular transactional workflow where calls to external agencies are normally required. The VPQ could

eliminate a significant portion of the wait time for employees making calls to other organizations and businesses and likewise increase productivity and gain financial benefits.

Purpose of the Study

The purpose of this research is to demonstrate through an experimental simulation study the functionality and operation of the VPQ within a call center. Specifically, will occupants within a VPQ experience a significantly reduced wait time. In addition, an exploratory research study will be conducted to determine consumer perceptions on waiting in service queues. Further investigation will be made into consumer perceptions of social justice that are aligned with Rawls' theory of social justice as that being justice as fairness (Rawls, 1971) What are consumer perceptions regarding social justice while waiting in a queue? Do customers feel a FIFO policy for service is the more socially just and equitable for all? How do customers perceive extended waits within a service queue? Do customers abandon queues because their patience time has been exceeded or do they abandon because they feel a social injustice has occurred?

Research Questions

In an attempt to understand the factors that contribute to queue abandonment and the potential for a VPQ to reduce wait times for select individuals in a queue, the following research questions were proposed for this study:

- Does patience threshold impact intent to abandon a service queue?
- Does perceived justice impact intent to abandon a service queue?
- During simulated operations, are call wait times statistically different between the standard service queue and the proposed VPQ?

The intent was to understand the factors that contribute to customers abandoning a service queue through an exploratory survey instrument and proposed theoretical framework. In the event

customer patience thresholds are a contributor to abandoning a service queue, a VPQ implementation can potentially alleviate that factor in some customers' decision to abandon a service queue. A DES was conducted to gauge the effectiveness of the proposed VPQ for creating advanced reservations to allow a select population access to faster service.

Summary of Appropriate Methods

According to Sekaran and Bougie (2013) a computer simulation is the most appropriate method to demonstrate the effectiveness of the VPQ in answering the proposed research questions. They state computer-based simulations are popular in business research due to their ability to reflect the effects of change in a system (Sekaran & Bougie, 2013). A quantitative study done through a survey instrument is appropriate to investigate the research questions around call queue abandonment. Having an available instrument or modifying an existing one is the preferred approach but if that is not an option, designing a specific instrument can be accomplished (Creswell, 2012). Question construction is a main challenge when creating a new instrument and specific guidelines are established to aid in formulating appropriate interrogatives (Babbie, 1990; Creswell, 2012)

Contribution to the Field of Business

The introduction of the VPQ and its effectiveness will undoubtedly offer resolution to business challenges in providing faster service to a select population. Businesses will have the opportunity to determine which of its valued customers may be served faster in an attempt to maintain and strengthen their relationship. Businesses can also utilize the VPQ to provide faster service to a select customer population. The VPQ can be deployed in various configurations to benefit both customers and businesses alike. A VPQ service could also provide standalone value for those who agree to a subscription or fee for service business model.

Definition of terms

Asymptotic Analysis: A method of describing limiting or bounded behavior (Borst et al., 2004)

Deterministic: No random behavior in determining future states of a system (Brown et al., 2005)

Discrete Event Simulation (DES): A model of a Stochastic process as it evolves over time by representing state variable changes at discrete points (Winston, 1994)

Erlang: A dimensionless unit that measures traffic in a communication system (Brockmeyer, Halstrom, & Jensen, 1948)

Exponential Distribution: The probability distribution that describes the time between events in a Poisson process (Koole & Mandelbaum, 2002)

Load Balancing: Distribute workload among available servers. When Poisson principles are violated and unpredictable behavior occurs, the service load must be balanced in order to maintain system integrity (Koole & Mandelbaum, 2002)

Markov Process: Process whose future behavior cannot be accurately predicted from past behavior and which involves random chance or probability (Newell, 1982).

Queuing Theory: Study of waiting lines (Zukerman, 2008).

Poisson Process: A counting process that has stationary increments if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval (Brown et al., 2005).

Steady State: Markov Chain transition matrix where probabilities have stabilized and behavior remains the same (Sankaranarayanan, 2011)

Stochastic Process: A random process evolving with time (Brown et al., 2005)

Limitations of Study

The survey instrument designed to investigate a person's intent to abandon a service queue was limited to a crowdsourcing approach with an Amazon Mechanical Turk (MTurk) deployment. Amazon's MTurk is an online marketplace where online workers complete Human Intelligence Tasks (HITs) such as responding to a survey. MTurk workers are compensated for their time in taking the survey and the incentive was not so large as to create an ethical concern (Creswell, 2012).

An appropriate survey instrument with items associated with the constructs under investigation was not available. The survey used during this investigation was constructed as an exploratory design with possibilities for future refinement and use. An attempt was made to ensure reliability of the developed instrument which included construct reliability with convergent and discriminant validity.

This DES portion of this study was limited to the M/M/1 single-server model for assessing the capability of the VPQ for delivering significant reductions in wait times for a select population of callers. Additional limitations are that the proposed experimental design does not account for call blocking before entering the queue or caller abandonment once in the queue since it was based on the Erlang-C model.

Literature Review

The basic fact about human existence is not that it is a tragedy, but that it is a bore. It is not so much a war as an endless standing in line.

-H. L. Mencken, A Mencken Chrestomathy

Definition of a Queue

A queue, or waiting line, describes a phenomenon where people are delayed in a line along with others waiting for a particular service. Taylor (1994) defines waiting for service as the time a customer is ready for service until the time service is provided. Some examples of queues include waiting in line at traffic lights for the light to turn green, waiting in a drive-through at a fast food restaurant, and waiting in line at a bank to deposit a check. Queues develop when the arrival rate exceeds the service rate and every arrival must be served (Gans et al., 2003). The study of waiting lines is called Queuing Theory and is considered a branch of operations research and management science disciplines (D. R. Anderson et al., 2013). The quantitative portion of the study of queuing theory borrows from many mathematical areas which include discrete math, probability theory, statistics, linear programming, and matrix operations (Winston, 1994).

Queues as a Social Phenomenon

Sorokin (1988) constructs a satirical dystopian reality of the Soviet citizen's life standing in never-ending service lines in *The Queue*. Written in dialogue style, the novel advances the story plot through conversation, a cacophony of random voices and whole-paragraph caesura with the intent to embed the reader in the story. In the introduction of the translated version, Sally Laird characterizes the Soviet propensity for queuing by stating "the basic principle of Soviet queuing is that you join the line first and *then* ask what it's for [original emphasis]" (Sorokin, 1988, p. i). Sorokin (1988) asserts that for the Soviet citizen, queuing is a natural act of

a conditioned society dependent on a state economy. The willingness to queue indicates an obedient subject and residents maintain a patriotic respect for being allowed to wait in line with fellow citizens. *The Queue* offers a glimpse of a lifestyle in which many individuals might not be able to relate. However, it does represent a unique aspect of a society burdened by limited resources and is complete with conflict, expectations, happiness, and love.

Additional evidence of queuing as a social phenomenon and human behavior within queues can be found in Andrews' (2013) *Why Does the Other Line Always Move Faster?* He delivers a humorous assessment of queues and how they are ubiquitous in everyday life. A cultural comparison is made between queue practices of western civilization and those of the former Soviet Union and Eastern Bloc nations. As alluded to in Sorokin's *The Queue*, the act of queuing by the populace indicates deference to authority and a compliant citizen. Evidence of cultural differences exist in many forms and is quite interesting. A similar comparison is mentioned regarding the British and their unwavering respect for queuing. One BBC report, complete with photographic evidence in Figure 7, documents a cash machine dispensing extra currency with citizens forming a queue to receive the cash. Queuing in their minds appears to be an inviolate principle that is followed, even when engaged in an ethically questionable activity. The orderly arrangement of the line suggests that at some level, queue integrity is maintained through a collective effort that demands an orderly flow through a process, irrespective of the reason or desire of the individual.

Queues as a social phenomenon can occasionally be modified according to certain cultural norms or expectations (Allon & Hanany, 2012). Andrews (2013) also discusses cultural differences in China regarding queuing.

Figure 7. Photographer Unknown. (2012). Retrieved November 15, 2106 from
http://www.bbc.com/news/uk-scotland-glasgow-west-20377784:_64207593_cash.jpg
Copyright 2012 by BBC. Reprinted with permission.

In preparation for the 2008 Summer Olympics, The People's Republic of China initiated campaigns to encourage their citizens to form queues, since they are practically non-existent as a societal norm. He points out that businesses have been able to monetize queuing by offering faster service to those willing to pay a premium (Andrews, 2013). Andrews (2013) also makes the assertion that a line-standing business has emerged where individuals can enter a monetary transaction for the ability to be served faster. He suggests that businesses are capitalizing on their own bad service by implementing various schemes whereby customers can pay to receive quicker service (Andrews, 2013). Closing out his interesting exposition on queuing, Andrews suggests our only defense is that we "become a systems person" (Andrews, 2013, p. 197) with deliberate consideration for engaging in activities that will contain some sort of queue "according to the probable behavior of others" (Andrews, 2013, p. 197).

Other cultures have modified the act of waiting in line into interesting behaviors that seem to ease the burden of standing in line, while maintaining a concept of fairness in the order in which individuals are served. In Figure 8, a surrogate (in this case, a pair of shoes) indicates individual position where maximum comfort and fairness is maintained while waiting in a queue.

Figure 8. "Queue: Thai Level" Photographer Unknown. (2012). Retrieved November 15, 2106 from http://i.imgur.com/eKgAP6O.jpg. *Used under Creative Commons Attribution 2.0 Generic license.*

Brady (2002) provides more thoughts on the social phenomenon emanating from queues. His study focused on the lines that developed in May 1999 while waiting for the premier of the newest Star Wars movie, *The Phantom Menace*. Brady (2002) initially intended to collect incidental information about the behavior of individuals occupying long service lines with a primary interest being the phenomenon of line segmentation and unprompted development of

cooperative groups of individuals. Brady's survey instrument was launched through a popular website associated with *The Phantom Menace*. The primary focus of the survey was to find out as much as possible about events occurring in the queue to purchase movie tickets. An attempt was made to determine size of each queue, duration of wait, location of each queue, established rules by each theater selling tickets, who exhibited leadership within the queue, who enforced queue discipline and what occurred when individuals attempted to cut into the line. However, he also observed the unanticipated behavior of "pre-scalping (Brady, 2002, p. 158) whereby people in the ticket line purchased more than a single ticket to resell for an economic advantage. He proposes queues are a microcosm of society where the study certain ethical behaviors can occur. Brady (2002) asserts queues provide the stage for the "quintessential ethical conflict" (Brady, 2002, p. 157) where individuals compete to further self-interests within a generally recognized framework of fairness with self-interest and civility observed within an environment of uncertainty and stress (Brady, 2002, p. 157). Brady acknowledges that an FCFS discipline seems to stand out as the prevalent queue organizing behavior, but circumstantial modifications can occur. He also maintains that FCFS can reduce the perceived stress associated with waiting in line because individuals are aware of their position in the queue and expected time remaining. In addition to gathering data on the general inquiry items, the research indicated that in several instances, line segmentation, or the formation of cooperative groups, enabled modifications to the FCFS principle Brady (2002).

The available literature on queuing theory is comprehensive and research within the last decade has provided extensive understanding of call center operations from a queuing science perspective (Alexander, MacLaren, O'Gorman, & White, 2012; Allon & Hanany, 2012; Franklin, 2010; Gans et al., 2010; Ibrahim & L'Ecuyer, 2012; Jouini, Aksin, & Dallery, 2011; Li

& Yue, 2015; Mandelbaum & Zeltyn, 2009; Mandelbaum & Momčilović, 2015; Mathew & Nambiar, 2013; Sankaranarayanan, 2011; Shen, 2010). The most recognized researcher and originator of queue studies was the Danish mathematician A. K. Erlang. Erlang had an affinity for mathematics at an early age and eventually found himself working for the Copenhagen Telephone Company at the beginning of the twentieth century (Brockmeyer et al., 1948). Erlang published his first work in 1909 titled “The Theory of Probabilities and Telephone Conversations” where he proved arriving calls followed a Poisson distribution (Brockmeyer et al., 1948; Erlang, 1909). In 1917, he published his most substantial work on queueing theory titled “The Number of Selectors in Automatic Telephone Exchanges” which laid the foundation for fundamental theories of telephone traffic even today (Brockmeyer et al., 1948). The majority of Erlang’s work was published in Danish journals and were translated into French by researchers who learned Danish in order to read his work in the original language (Brockmeyer et al., 1948). Erlang’s contributions were so significant that a unit of measure for telephony offered load was named in his honor. The *Erlang* (E) is a dimensionless unit that measures the offered, or carried load, on a telephone circuit. Another important concept named after Erlang is the *Erlang distribution*. The Erlang distribution models a continuous random variable whose density function is specified by a rate parameter R and a shape parameter *k* where *k* is the set of all positive integers (Winston, 1994). The Erlang distribution is appropriate when interarrival times to a queue do not follow an exponential distribution and for increasing values of *k*, the Erlang distribution exhibits characteristics of a normal distribution. The Erlang distribution probability density function is represented by

$$f(x;k,\lambda) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \; for \; x, \lambda \; \geq 0. \qquad (5)$$

At large values of *k*, the Erlang distribution approaches a random variable that has zero variance representing a constant interarrival time (Winston, 1994). The Erlang distribution is more flexible in modeling interarrival times due to the shape parameter *k*. Other significant queuing science concepts attributed to Erlang are the descriptions of queueing models representing various arrival and service distributions.

Major Research on Queuing Theory

Call center queue studies have included a broad scope of investigation primarily in queue arrival and queue service investigations. Ibrahim and L'Ecuyer (2012) compared different statistical models for call arrivals through evaluation of forecast accuracy based on varying lead-times which ranged from hours to weeks in advance. The results show a bivariate model has the best potential to predict arrivals and should be investigated further. Lewis, Herbert, Summons, and Chivers (2007) proposed an agent-based simulation of a multi-queue emergency services call center to discover efficiencies in service agent staffing to handle unexpected increases in call volumes. Mandelbaum and Zeltyn (2007) assert the Erlang-A model is more appropriate to model call center arrivals since it accounts for an exponentially distributed customer patience time θ. Massey (2002) demonstrates the suitability of time-varying rate queues for telecommunications models. Paek and Horvitz (2004) investigated the use of IVRs for call arrivals and developed a predictive model on when call failure might occur in order to transfer the caller to a human service agent. Their study included eight factors on possible call outcomes from which they developed predictive models on when to transfer calls. Robbins, Medeiros, and Harrison (2010) question the utility of the Erlang-C model in call centers that experience high utilization rates. Their assertion is that the Erlang-C model produces large errors in this situation due to higher utilization rates, caller abandonments and fewer service agents. Shen (2010)

emphasizes a need for more investigation regarding call forecasting to support customer service agent staffing and to effectively model customer retrial behavior and what portion of arrivals are actually callers who previously abandoned or exceeded their patience time. Wallace and Whitt (2005) discovered limited cross-training of service agents produced results similar to scenarios where agents had all skills necessary to provide service to arriving customers. Weinberg, Brown, and Stroud (2007) investigated a Bayesian approach that resulted in development of point estimates and complete distributions of variables of within-day arrivals of calls to a U.S. commercial bank call center. They maintain this forecasting methodology was superior to previous model research conducted by Brown et al. (2005) who employed a least squares fit procedure.

Queue service research is as comprehensive as arrival studies and generally seeks to hasten service to the queue population at a desired service level to avoid abandonments, reduce costs and produce a better customer service experience (Borst et al., 2004; Gans et al., 2003). Atlason, Epelman, and Henderson (2004) conducted research on a simulation and cutting plane method to minimize service costs over multiple time periods. The results of their investigation indicated the proposed method has potential to solve optimization problems where some constraints can only be investigated through simulation. Staffing costs of call centers historically consume the majority of the operating budget (Akhtar & Latif, 2010; Aksin et al., 2007; Borst et al., 2004; Weinberg et al., 2007). Borst et al. (2004) develop a framework for asymptotic optimization of a queue that balances service cost and desired service level. They propose three regimes of operation that comprise a quality-driven aspect, efficiency-driven approach and a rationalized regime that establishes an equilibrium between the former two. Recognizing that call centers should not necessarily be considered *cost centers* in some cases from a financial

perspective, Demiriz, Kula, and Akbilek (2009) propose a cross-selling framework to generate revenue without degrading service quality. Green, Kolesar, and Soares (2003) developed an enhanced heuristic approach to solving the staffing problem for call centers. Their research indicated the preferred method of determining staffing levels had limitations in determining predicted queue lengths, time lag between actual customer demand and system lag and planning period arrival rates. Gurumurthi and Benjaafar (2004) offer a modeling framework where different customer classes can be served by more than one type of server. The results of their research show there is considerable flexibility in control procedures to increase system throughput. The researchers also suggest there is a possibility of reconsidering the assumptions of Poisson arrivals and exponential service intervals through their model's flexibility characteristics. Mandelbaum and Momčilović (2015) are currently investigating queues where individual information is known prior to queue entry. They propose a many-server fluid model where customers with shorter estimated patience times are given priority over customers with longer patience estimates. The results are promising in that the proposed model shows a Least-Patience First (LPF) routing scheme is better than FIFO during periods where queue delays are approaching customer patience times. Whitt (2005) reiterates the conclusions of Mandelbaum and Zeltyn (2007) in stating that service and abandon time distributions are not exactly exponential. He proposes a new state-dependent, Markovian, steady-state approximation model $M/GI/s/r+GI$ which has a Poisson arrival process, independent and identically distributed (IID) service times with a general distribution, s servers, r extra waiting spaces and IID customer abandonment times with a general distribution. Queue models of this form are extremely difficult to analyze due to the fact that normal birth-death steady-state equations are no longer valid since

the probability of service now depends on the length of the time interval between the last service completion and the Markovian memoryless property is violated (Winston, 1994).

Applications within Call Centers

Call centers are an increasingly important element in today's business operations and are normally considered large service entities where customer service agents provide a variety of services over the telephone (Gans et al., 2003; Shen, 2010; Weinberg et al., 2007). Even with the ubiquity of the internet and electronic global communication ability, the telephone is expected to remain a primary contact channel and be the preferred method for customer complex service requirements (Deloitte, 2015; Gans et al., 2003; Koole & Mandelbaum, 2002; Mathew & Nambiar, 2013; Weinberg et al., 2007). The natural successor to the call center is the contact center where all forms of customer contact are handled (Gans et al., 2003; Koole & Mandelbaum, 2002). In addition to providing service over the phone, a contact center incorporates other multi-media customer contact channels such as e-mail, fax, internet chat and IVR units. IVRs allow customer self-service for many transactions via telephone keypad entries or voice responses (Gans et al., 2003). Virtualization now enables seamless operations between contact centers in different geographic locations (Gans et al., 2003). Customer initiated calls are considered *inbound* calls and are predominant during normal business operations. Call centers also conduct outbound calls to customers for sales or marketing purposes (Aksin et al., 2007; Ali, 2010; Holman et al., 2007; Mehrotra & Fama, 2003; Shen, 2010).

Markov and Birth-Death Processes

Queues develop when arrival rate for service exceeds the service rate of the servicer and the system state can approach infinity if this were to continue indefinitely. Queues are represented by Kendall notation, which classifies each in terms of arrival process, service time

distribution and number of servers (Winston, 1994). A simple representation of a queue is illustrated in Figure 9 showing a typical M/M/1.

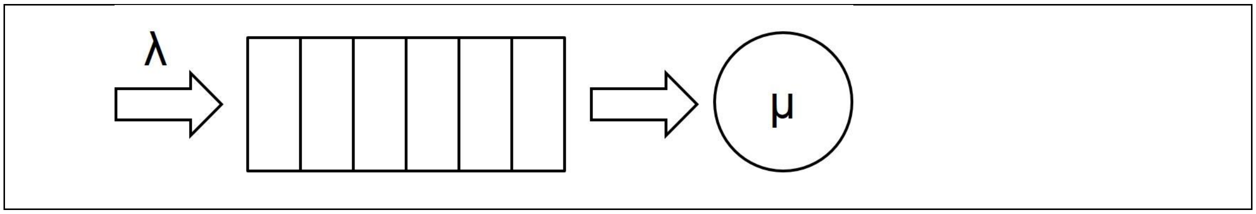


Figure 9. M/M/1 Queueing Model.

This diagram indicates arrivals entering the queue follow a Poisson process with average arrival rate indicated by λ, a service time distribution that is Markovian as well and a single server providing service at rate μ. Since the Poisson process is a discrete random variable distribution describing occurrences over a specific time interval, the arrivals are independent of one another within the queue (D. R. Anderson et al., 2013). The probability mass function for a Poisson process is indicated by

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \; for \; x = 0, 1, 2, 3 \ldots \tag{6}$$

where *x* is the number of events being predicted and λ is the expected value, or mean number of occurrences in the interval. The M/M/1 queue is a stochastic process with a state space being the set (0, 1, 2, 3…) where the value equals the number of customers in the system, including customers receiving service. It is also a special type of stochastic process called a *Markov Process*, or *Continuous Time Markov Chain* on the non-negative integers with transitions from state *i* to *i* + 1 at rate λ. A Markov process is said to be memoryless in that the future state of the system depends solely on the present state and not any of the preceding states of the system. The state transition diagram is shown in Figure 10.

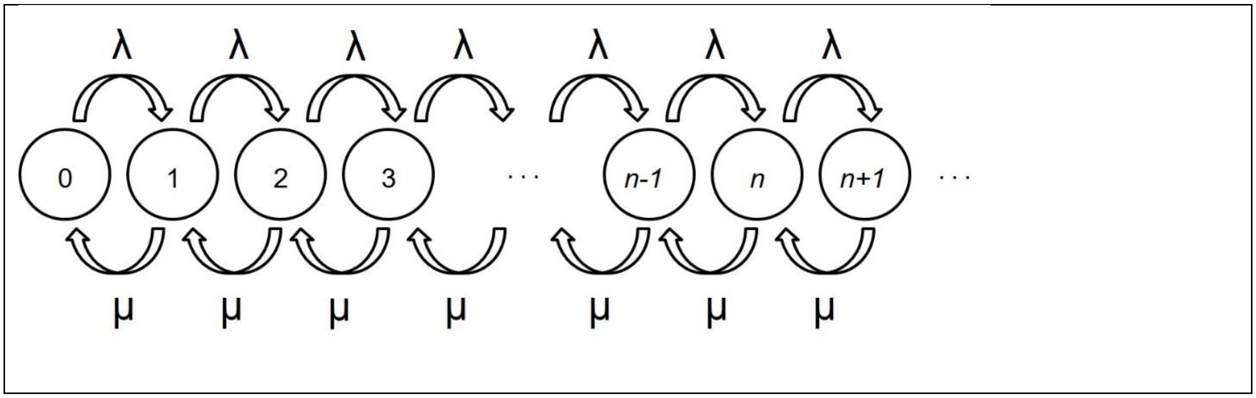


Figure 10. State transition diagram for M/M/1 queue.

Queuing systems that exhibit exponential interarrival and service times are also considered a generalized *birth-death process* where the probability of moving from one state *i* to *i* + 1 (a *birth*) is equal to the probability of going from *i* + 1 to *i* (a *death*) and is indicated by

$$P_{i+1}(death_{i+1}) = P_i(birth_i) \tag{7}$$

where the probability of a death in state *i* + 1 is equal to the probability of a birth in state *i*. The ratio of the arrival rate λ to service rate μ is termed λ/μ and is called the *utilization factor*. This represents the probability that an arriving customer must wait for service because the server is busy or being utilized (D. R. Anderson et al., 2013). The probability of the system being in state 0 with no customers in service or the queue is given by

$$P_0 = 1 - \lambda/\mu \ . \tag{8}$$

The probability of the system being in any state is given by

$$P_{i+1} = (\lambda/\mu)P_i \text{ for all } i \ . \tag{9}$$

Therefore, the probability for the system being in any state *i* is given by the equation

$$P_i = (\lambda/\mu)^i P_0 \ . \tag{10}$$

Summing all the system state probabilities will equal to 1 and is represented by

$$\sum_{i=0}^{\infty} (\lambda/\mu)^i \, P_0 = \sum_{i=0}^{\infty} P_i = 1\,. \tag{11}$$

From calculus, the convergence of an infinite series of this form is given by

$$\sum_{x=0}^{\infty} x^i = \frac{1}{1-x} \; for \; 0 < x < 1\,. \tag{12}$$

Substituting the term λ/μ from equation (6) into this equation gives

$$\sum_{i=0}^{\infty} (\lambda/\mu)^i \, P_0 = \frac{1}{1-(\lambda/\mu)} \, P_0 = 1\,. \tag{13}$$

Replacing λ/μ with ρ gives the alternate form

$$\sum_{i=0}^{\infty} \rho^i \, P_0 = \frac{1}{1-(\rho)} \, P_0 = 1\,. \tag{14}$$

The expected value of the number of units in the system can be calculated using the formula

$$E = \sum_{i=0}^{\infty} iP_i = \sum_{i=0}^{\infty} i\,(\lambda/\mu)^i P_0 = \sum_{i=0}^{\infty} i\rho^i \, P_0\,. \tag{15}$$

This equation can be factored into the following form

$$\sum_{i=0}^{\infty} i\rho^i \, P_0 = \sum_{i=0}^{\infty} i(\rho)\left(\rho^{i-1}\right)P_0 = (1-\rho)(\rho)\sum_{i=0}^{\infty} i(\rho^{i-1})\,. \tag{16}$$

From calculus, the convergence of an infinite series in this form is given by

$$\sum_{i=0}^{\infty} ix^{i-1} = \frac{1}{(1-x)^2} \; for \; 0 < x < 1\,. \tag{17}$$

Therefore, the expected value of the number units in the system is

$$(1-\rho)(\rho)\sum_{i=0}^{\infty} i(\rho^{i-1}) = \frac{(1-\rho)(\rho)}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}\,. \tag{18}$$

The transition rate matrix for an M/M/1 queue is represented below

$$Q = \begin{pmatrix} -\lambda & \lambda & & \\ \mu & -(\mu+\lambda) & \lambda & \\ & \mu & -(\mu+\lambda) & \lambda \\ & & \mu & \dots \end{pmatrix}. \tag{19}$$

Little's Law states the average number of units L in a system operating at a steady-state can also be determined by multiplying the arrival rate λ by the average time spent in the system W and the generally accepted form of the equation is

$$L = \lambda W\ . \tag{20}$$

Little's Law applies to any queue model regardless of whether arrivals follow a Poisson probability distribution or service times follow an exponential distribution (D. R. Anderson et al., 2013).

Erlang Models

The M/M/1 model is also referred to as the Erlang-C and it is commonly used to assess call center operational performance because of its simplicity with respect to basic assumptions and ease of use (Brown et al., 2005; Gans et al., 2003; Garnett et al., 2002; Mandelbaum & Zeltyn, 2009; Robbins et al., 2010). However, the appropriateness of the Erlang-C to model call center operations has been called into question by some researchers who insist its simplistic nature under steady-state conditions does not account for all call center variables, such as customer patience, call blocking or abandonments (Brown et al., 2005; Gans et al., 2003; Garnett et al., 2002; Koole & Mandelbaum, 2002; Mandelbaum & Zeltyn, 2009; Mandelbaum & Zeltyn, 2007; Robbins et al., 2010; Shen, 2010). The Erlang-A is a variation of the Erlang-C and is considered the preferred model for capturing customer abandonments at rate θ (Mandelbaum & Zeltyn, 2007). The Erlang-A is generally written in the form M/M/*n*+G where G represents the Markovian distribution of the customer's patience time (Brown et al., 2005; Mandelbaum & Zeltyn, 2009; Shen, 2010). The final model is the Erlang-B, which is also referred to as *Erlang's*

loss formula and represents a system where calls are blocked and usually cleared when all available servers are busy and is generally presented in the form *M/M/n/n* which incorporates the number parallel servers and maximum number of customers in the system (Winston, 1994).

Economic Analysis of Waiting Lines

Determining economic costs of waiting from a business perspective is relatively easy to assess given that certain assumptions, or desired service objectives, are provided (D. R. Anderson et al., 2013). A total cost *TC* model can be constructed using the following equation

$$TC = c_w L + c_s k \tag{21}$$

where c_w is the waiting cost per customer per unit time and c_s is the service cost for each server in the same time period as depicted in Figure 11.

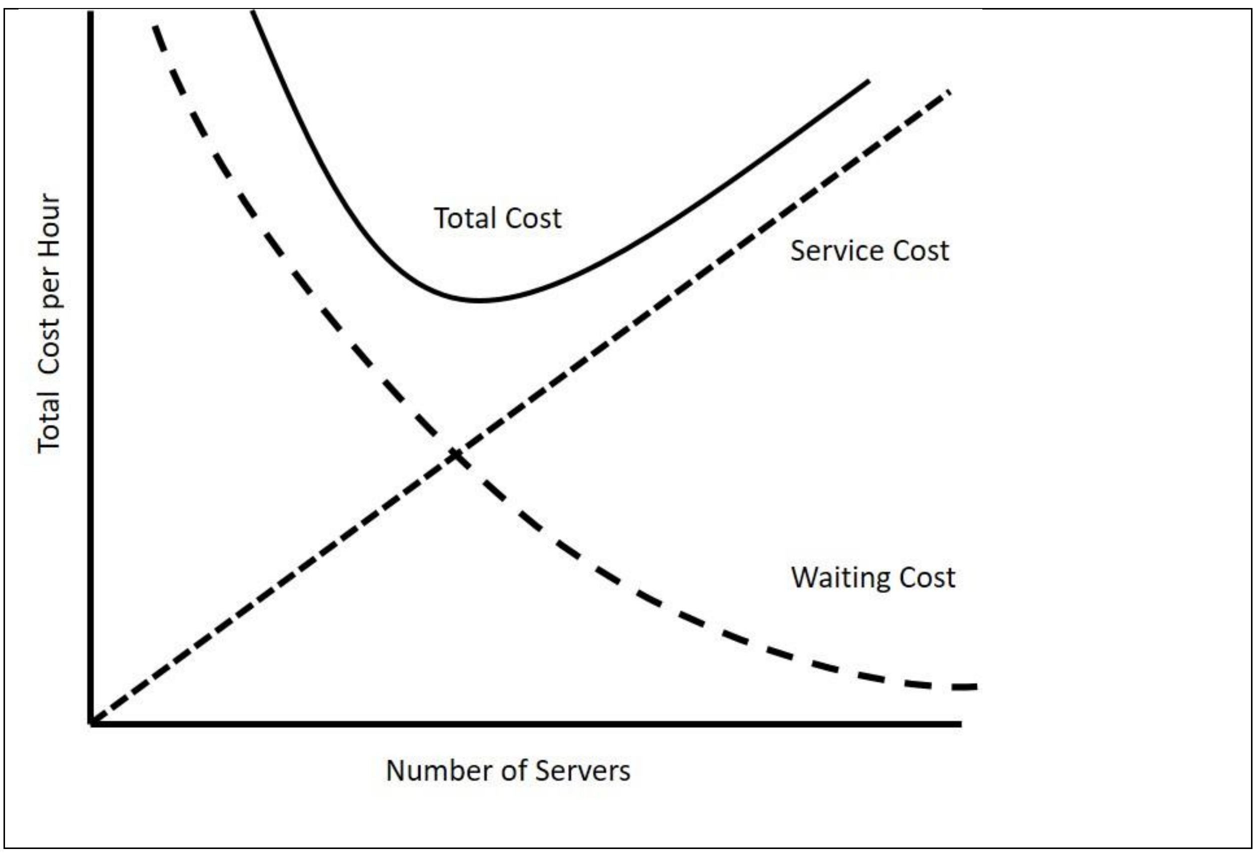


Figure 11. Total Cost curve as a function of service and waiting cost.

The term L is the average number of customers in the system and k is the number of servers. The assumptions for waiting cost are not direct costs to the servicer but are related to the fact that a customer who decides not to wait, or abandon a queue, when wait times are lengthy equates to potential loss of revenue and therefore a cost to business (D. R. Anderson et al., 2013).

Queue Disciplines

The order in which customers are served is generally regarded as the queue *discipline* and other schemes, in addition to FIFO, are last in, first out (LIFO), service in random order (SIRO) and priority service based on certain customer characteristics (Zukerman, 2008). Although FIFO is generally recognized as the predominant queue discipline, other systems have been investigated and provided interesting conclusions about their performance. Mandelbaum and Zeltyn (2007) maintain that when organizations do not take into account customer abandonments and focus only on reducing the average wait time of those in the queue, a LIFO policy would optimize their service metric, but would also produce a negative experience for customers who joined the queue at an earlier time. Larson (1987) uses a case study from the Boston Police Department from the 1960s to illustrate the point that calls answered in random order have the same mean time spent waiting in the queue. Early telephony technology only allowed the police department operators to select one of the blinking lights on their phone and many times when calls were waiting, the operator did not remember the order in which the calls arrived and randomly answered calls waiting on hold (Larson, 1987). Priority queues have been implemented in many service industries and allow a select population to receive service faster than those waiting in the regular queue but they can also cause perceptions of social injustice in certain settings (Alexander et al., 2012; Andrews, 2013; Brady, 2002).

Consumer Psychology in Queues

The psychology of waiting in line has received considerable research in formulating consumer perceptions of service quality and experience (Larson, 1987; Shen, 2010). Is FIFO the *de facto* standard for queues from a social phenomenon perspective? Individuals standing in a service line exhibit a perfect example of the fundamental geometry of a queue through its linear congruence where one side is for entry and the other leads to desired service. The majority of queue studies presumes FIFO is the socially just queue discipline and violations of that principle can lead to conflict (Larson, 1987; Maister, 2005). Alexander et al. (2012) discovered that social justice and equity can collide when priority queues that allow some participants to pay for faster service are created. The researchers used a mixed methods approach combining observations and personal interviews with follow-on experiments to explore the effect of priority queues on satisfaction, word of mouth and purchase intentions. Their 2 X 2 between-subjects factorial design included two levels of wait times with the second factor being the absence or presence of a priority queue. The results of their experiments indicated that an increased queue length had a negative effect on customer satisfaction and purchase intent. Additionally, longer wait times had a negative effect on customers in the main queue and a positive effect on customers in the priority queue. Alexander et al. (2012) conclude that when social justice is achieved for those in the main queue, then an inequity exists for those who have paid for priority service. Likewise, when equity is created for those in the priority queue, an injustice occurs to those in the main queue.

Larson (1987, 1988) explored social justice and the psychology of queuing and discovered that other factors are involved in a person's perceived value of waiting in line. Larson relates his personal experience of receiving a bicycle he purchased for his son at a department

store pickup window. After a lengthy waiting period, he notices others who have arrived after him received their purchases first. Larson also asserts that adherence to an FCFS policy is perhaps the universal discipline for creating social justice within a queue by stating “Queueing theorists and social scientists have long believed that first-come, first-served (FCFS) is the socially just *queue* discipline and first-in, first-out (FIFO) the socially just *system* discipline [original emphasis]” (Larson, 1988, p. 896). Larson also discussed the concept of *slips* and *skips* and how they can circumvent FCFS and lead to the breakdown of social justice for a queue. A slip is when an individual joins a queue but is served before someone who entered the queue previously. A skip occurs when an individual is advanced in the queue and is served before others who entered before them. A person who experiences a slip has been wronged but one who can skip has been given preference in some way over others who arrived in the queue before them. Larson describes situations where someone’s fear of social injustice impacts their perceived value in waiting in a queue through customer satisfaction survey results from a management science consultant on single queue operations in Wendy’s, McDonald’s, and Burger King restaurants. In some Wendy’s locations, customers preferred the single queue with almost double the wait time to the multi-queue operations of McDonald’s and Burger King where social injustice may occur due to later arrivals being served first. Social justice implications and perceived wait time within queues are influenced by the queue environment and the amount of *empty time* created (Larson, 1987). The example Larson uses to illustrate this point is a bank in Cambridge, Massachusetts that uses live entertainment in the form of music, exhibits and shows during heavy traffic hours (Larson, 1987). These activities, although they do not offer any efficiencies for faster service, eliminate empty time and the perceived wait time seem less lengthy. Larson also recognizes the importance of feedback to persons waiting in a queue. Based

on his own personal observations, Larson (1988) states that customers are more at ease when they have some idea in advance of their expected wait time. It is this context where Larson (1988) introduces an observation that in some cities with 911 emergency service, the police departments deliberately delay lower priority calls for service in order to maintain the ability to effectively respond to more severe situations. This is an example of a priority queue discipline where judgment is made by the servicer (police department) on who gets service faster. This priority scheme can and likely will violate a FIFO policy. But according to Larson (1988), citizens who are told of an expected delay for lower priority calls are more satisfied when the police arrive within the stated time period as opposed to those with no knowledge of expected arrival, even if the wait time was shorter. Larson (1988) draws the conclusion that other factors affect a person's perception and utility of the wait within a queue by stating "Queuing theorists are starting to realize that what happens to you while in line is more important than how long you're there" (Larson, 1987, p. 61). His position is that further investigation of queues should focus beyond the traditional quantifiable variables associated with Little's Law.

Maister (2005) also explored the psychology of waiting lines and provided anecdotal evidence from personal observations that humans do value fairness and social justice when faced with waiting in a queue. His major premise is that our waiting line experience directly impacts our perceived quality of service. Maister (2005) elaborated on the waiting experience by offering eight propositions on the psychology of waiting. The first proposition is that "Occupied Time Feels Shorter Than Unoccupied Time" (Maister, 2005, p. 3). The understanding of this statement stems from an individual's perception of time passage. If attention is given directly to time spent in a queue, as well as the anticipatory anxiety arising from not knowing how long of a wait remains, one may feel an overwhelming perception of a lengthy delay in queue. He provides as

an example the approach some restaurants take in allowing customers to delay in the bar while they wait for a table. The next proposition is “People Want to Get Started” (Maister, 2005, p. 3). For those waiting in a queue, the anxiety can be reduced by the fact that they know that the service has begun and this provides a temporal break between waiting in line and being served. Again, the relevant example of this is in the restaurant industry where the wait staff may acknowledge a party that was just seated and offer menus. A physician’s office may respond immediately to new patient arrivals by taking their vital signs before seeing the doctor. Maister’s third proposition is that “Anxiety Makes Waits Seem Longer” (Maister, 2005, p. 4). An example of this would be worrying about the line chosen at the supermarket or waiting to board an airplane. Both situations involve giving up control to the servicer with the potential of some injustice occurring once that control is relinquished. Maister’s forth proposition is “Uncertain Waits Are Longer than Known, Finite Waits” (Maister, 2005, p. 5). Maister (2005) states that the most severe form of anxiety can occur when the wait time is unknown. When given a reliable and accurate estimate of the wait time, individuals feel more at ease, even though the wait estimate may be quite lengthy. According to him, “The wait until the appointed time is finite; waiting beyond the point has no knowable limit” (Maister, 2005, p. 5). His fifth proposition is “Unexplained Waits Are Longer than Explained Waits” (Maister, 2005, p. 5). Our expectations for waiting are influenced by our past experiences and known environmental conditions (e.g., flight delays due to weather). When the wait time exceeds our internal estimate based on the known conditions, waiting beyond that time is unexplained and leaves a feeling of uncertainty. Regardless of the reason for the delay, Maister emphasizes that any information for the delay is more reassuring than no explanation at all. Maister’s sixth proposition states that “Unfair Waits Are Longer Than Equitable Waits” (Maister, 2005, p. 6). This statement refers to an injustice

occurring when the expectation and observation within a queue should follow a FIFO scheme. An example of this is when a customer witnesses another party being seated in a restaurant ahead of them. Strict enforcement of the FIFO policy can occur when servicers provide numbers for service and there is no allowance for priority service. Maister's seventh proposition is "The More Valuable the Service, the Longer the Customer Will Wait" (Maister, 2005, p. 7). This proposition explains why the tolerance for waiting depends on the perceived value of the service. For example, Maister offers the example that someone with a full grocery cart will likely tolerate a longer checkout line. Customers with fewer items resented the fact they had to wait in long lines, which led many grocery stores to offer an express lane for faster service. Maister's final proposition states that "Solo Waits Feel Longer Than Group Waits" (Maister, 2005, p. 8). He proposes that waiting as a group lends itself to developing a sense of community and waiting in a line itself offers a form of sociological value much like that found in Brady's (2002) research on queue behavior. Maister (2005) declares that his eight propositions are by no means all-inclusive of psychological factors experienced in queues, but they can be used as starting points for future research on the contextual factors affecting customer satisfaction.

Customer Expectations

Some research on customer service expectations has focused on understanding the nature and determinants of those expectations within a general service setting (Zeithaml, Berry, & Parasuraman, 1993). Zeithaml et al. (1993) proposed a model specifying three different types of service expectations which include desired service, adequate service, and predicted service. Their exploratory research provided a theoretical framework which included seventeen propositions and included antecedents for each of the theorized expectations. Zohar, Mandelbaum, and Shimkin (2002) investigated patience adaptation as a function of their service expectation of

anticipated wait time. They proposed an *M/M/m* model that considers adaptive customer behavior influenced by changes in anticipated wait time. The findings suggest the model to be applicable for a steady-state equilibrium system for determining patience based on average wait time. Other research has focused on the relationship between wait time and perceived service evaluations. (Taylor, 1994) focused on understanding how delays in service affected overall customer service evaluations. Three types of waiting is defined in this study and generally, *waiting for service* is considered the time from when a customer is ready to receive service until the service begins (Taylor, 1994). The three types of waits include a pre-process wait, an in-process wait and a post-process wait (Taylor, 1994). A pre-process wait would be considered waiting for a table at a restaurant. An in-process wait would be the wait between meal ordering and service and a post-process wait would be considered waiting for and paying the bill after the meal was complete (Taylor, 1994). Maister (2005) suggests pre-process waits are experienced by customers differently than in-process and post-process waits. Taylor's (1994) study focused only on a pre-process delay for service and reference to previous studies suggesting a negative relationship between queue delays and overall evaluation of service contributed to a robust theoretical framework which contained constructs for punctuality, uncertainty and anger (Taylor, 1994). The findings suggest there is a negative relationship between longer pre-process delays and overall evaluation of service (Taylor, 1994). Interestingly, it was found that punctuality was influenced more by the anger created for the delay than the actual duration of the delay itself (Taylor, 1994). This finding supports previous suggestions that servicers should be more concerned with managing a person’s perception of waiting than the wait duration itself (Larson, 1987; Maister, 2005).

Impatient Callers and Abandonment

Callers who initially enter a queue but subsequently leave are said to abandon the queue or *renege* (Gans et al., 2003). Customers who do not enter the queue at all are said to *balk* when their impatience is greater than the perceived value of the service (Aksin et al., 2007). Mandelbaum and Shimkin (2000) investigated a model for rational queue abandonments with the assumption that callers have a given patience distribution for waiting in a queue. Their findings suggested there is a rational framework for determining the queue abandonment period given that customers have no information on their position in the queue. Several queue studies have focused on abandonment as a result of customer patience (Shen, 2010) with consideration for time spent waiting, or what some researchers consider the “sunk cost effect” (Zhou & Soman, 2003). Zhou and Soman (2003) researched the effect of the number of people behind in a queue on consumer psychology and the likelihood of reneging. They suggest that in addition to the number of people ahead in a queue, the number of individuals waiting in line behind is a key factor in the decision to renege. The researchers theorize that consumers make social comparisons of their position in a queue in a downward direction. Specifically, they suggest individuals assess their position in a queue and seeing more people behind them leads to a more positive self-assessed affective state of well-being. Although the wait itself may be intolerable to a certain degree, many may relate to looking in the rear view mirror in a long line of traffic and seeing quite a few less fortunate people behind them. Zhou and Soman (2003) theorize there are three factors associated with social comparisons made by individuals when subjected to the number behind effect: queue factors, individual factors and situational factors. Queue factors refer to the salient information about the queue from a consumer’s perspective. For example, a linear, well-defined queue where individuals are prominently ordered from beginning to end

conveys ample information on the nature and composition of the queue. A consumer's assessment of their relative position in such a queue would be easier to discern, as opposed to other systems where relative position is not as prominent, such as a ticketing system where consumers are assigned a number based on their arrival to the queue. Individual factors refer to the notion that certain types of individuals tend to make social comparisons more than others. Gibbons and Bunk created and validated the Iowa-Netherlands Comparison Orientation Measure (INCOM) which measures the individual differences in the tendency to make social comparisons (Gibbons and Bunk, as cited in Zhou & Soman, 2003). Individuals scoring higher on the INCOM scale had a greater tendency to make social comparisons and the expectation was those individuals would also demonstrate a stronger number behind effect (Zhou & Soman, 2003). According to the researchers, situational factors are those circumstances where social comparisons to others who are behind in a queue can occur. They cite one such influence to be the notion of counterfactual thinking, Counterfactual thinking is when individuals ascribe meaning to fortuitous circumstances which can lead them to believe they are lucky or in a higher affective state than normal. Zhou and Soman (2003) assert that situational factors which lead to counterfactual thinking will influence those individuals with the tendency for downward social comparison and increase their likelihood of engaging in that behavior. The results of their investigation show that the likelihood of reneging is less when the number of people behind in a queue is large. Additionally, the researchers validated their assumptions on the three proposed factors associated with social comparisons. They found that the effect of the number behind in a queue was stronger when relative position in a queue was known. Additionally, the factors involving the INCOM scale for social comparisons and tendency to engage in counterfactual thinking also produced similar results and verified the increased number behind effect for those

individuals. Zhou and Soman's (2003) study is truly unique in that it investigated the psychological effect of the number of people behind in a queue, as opposed to much research on positioning in a queue and expected costs or benefits of continued waiting. The results are largely conclusive in suggesting consumers naturally make downward comparisons within queues of those behind who are in a worse position than they occupy.

Abandonment in many situations can be attributed to understaffing and lead to retrials for those customers who abandoned the queue, or were blocked, on their first attempt (Aguir, Karaesmen, Aksin, & Chauvet, 2004; Gans et al., 2003; Mandelbaum & Shimkin, 2000; Shen, 2010). Aguir et al. (2004) investigated the impact of retrials on call center forecasting and staffing levels. They found that there is a significant relationship between staffing levels and customer retrials. Underestimating the demand due to retrials can subsequently lead to further distortions in anticipated service levels. Garnett et al. (2002) developed a set of rules regarding large call center design through asymptotic behavior analysis of performance in a rationalized regime. Guo (2007) studied delay information to customers that was provided in three forms: no information, partial information and exact information. Partial information was limited to providing total system occupancy and full information was the exact waiting time. The results of his study showed increasing the amount of information available can be detrimental to both customer and servicer, depending on the customer’s delay sensitivity distribution parameter. Jouini et al. (2011) investigated delay announcements on customer balking and reneging through three models that provide perfect delay information, no delay information and delay announcements respectively. The first two models were considered benchmark cases to rationalize customer behavior within the framework of the third model, which provided delay information at various coverage levels. The results show there is a definite relationship between

reneging and balking in the presence of delay information and more delay information to the customer is not necessarily better. Zohar et al. (2002) also proposed a dynamic learning model in their investigation of customer adaptive patience in the presence of changing wait times.

Queues and Social Justice

What is the basis for the universal applicability of FIFO for a servicing system? Queues in general are naturally assumed to follow a FIFO discipline, especially when a system maintains the generalized sociological phenomenon of an FCFS policy (Alexander et al., 2012; Brady, 2002; Larson, 1988). How are queues addressed in the context of the law and legal interpretations? Perry and Zarsky's (2013) legal research report published in the *Iowa Law Review* focuses in depth on the FIFO principle with regard to case law, legal interpretation and use as an “extra-legal norm” (Perry & Zarsky, 2013, p. 5). Within the context of the law, the authors consider FIFO an allocation method, particularly as it pertains to property law and the concept of a lien position on a collateral security. Perry and Zarsky's (2013) main purpose was to provide a theoretical framework for evaluating FIFO rules, exceptions to policy and practical applications. The authors also demonstrate and evaluate FIFO’s role in law through investigation of its pervasiveness in case law through implementations that are typically lacking of strict enforcement criteria or codified meaning. Perry and Zarsky's (2013) final intent was to demonstrate that FIFO is not a compact solution for every legal application. They propose that FIFO applicability should be assessed subjectively for each situation and use in one context does not necessarily mean a unbending categorical legal imperative (Perry & Zarsky, 2013). A discussion of fairness occurs at the beginning of the research paper which starts by acknowledging the widely-held belief by many scholars that FIFO is the socially just system discipline (Larson, as cited in Perry & Zarsky, 2013). The authors state their main purpose is to

provide a theoretical framework for evaluating FIFO rules, exceptions to its use and real-world applications. Their opening approach in the context of their study was to use the concept of fairness and efficiency as a starting point in developing a cogent argument for the legal basis of a FIFO system. Perry and Zarsky's (2013) secondary purpose was to demonstrate and assess FIFO's application within the law. The tertiary purpose of their research was to identify FIFO's principle strengths and weaknesses when applied in a legal context. In addressing *fairness*, the authors submit to previous scholarly interpretations that describe a FIFO system as a socially just and *ethical* system (Brady, as cited in Perry & Zarsky, 2013). In the preliminary discussion of fairness, the authors illustrate what it means to be *fair* in the context of a FIFO allocation scheme by stating that a fairness scheme "must be compatible with the most fundamental human perceptions" (Perry & Zarsky, 2013, p. 8). An even deeper connotation for societal support of a FIFO policy is indicated in the direct assertion that "perceptions of fairness matter because complying with or violating one's perception of fairness impinges on one's welfare" (Perry & Zarsky, 2013, p. 8). In discussing normative fairness and classless applications of a FIFO policy, Perry and Zarsky (2013) invoke the concept of Rawls' *veil of ignorance* (discussed in much more detail later) when rational individuals derive a fair and just system of allocation when unaware of individual talents, tastes and social standings. From an egalitarian defense of FIFO, the authors state that FIFO "assumes all allocation participants are roughly equal in all relevant respects" (Perry & Zarsky, 2013, p. 13). An alternative comparison is made between FIFO and what the author's call *random selection for service* when irrelevant criteria and indistinguishable characteristics are considered. They likewise agree with other investigators in stating that an allocation scheme between both methods would result in quantifiable measurements of average wait and service times being practically equal, as well as a blind indifference to personal and

social characteristics of participants (Larson, 1987; Perry & Zarsky, 2013). One weakness of FIFO offered by the authors is that FIFO can allow the more affluent or those with greater economic means to advance in a queue or compensate others willing to serve as a surrogate for economic advantage (Perry & Zarsky, 2013). A more detailed discussion of that situation will be discussed later. Consideration of the positive aspects of a FIFO policy reveal that it explicitly values everyone's time. Reflecting on the adage that *time is money*, the authors maintain that in some regard, FIFO credits a person's time spent in a queue with a theoretical monetary value. In addition, the waiting time differential experienced by each individual may allude to a level of value they pursue when occupying a queue. Individuals who are aware of increased variance in waiting times can choose to delay their entry or perhaps not enter the queue at all (Perry & Zarsky, 2013). The authors also frame the concept of a FIFO policy with the legal notion of desert. Desert claims are those which are justified by a party A which deserves an outcome B because of some basis C (Perry & Zarsky, 2013). For example, a student who studies well deserves a good grade on an exam. The authors also state several cases where FIFO would not hold for certain desert claims where level of effort does not warrant the same level of reward. Perry and Zarsky (2013) also delve into the efficiency aspects of FIFO as an allocation mechanism. Within that realm of thought, the authors propose both *ex post* and *ex ante* FIFO effects. *Ex post* FIFO applications are suggested in cases where employee seniority holds a higher advantage in job security or in property law where creditors occupy lien positions of real property collateral. Temporal positioning in this case reflects the subjective effort of a party in entering a queue to gain an equivalent reward. *Ex ante* effects deal with a party's conduct before entering a queue and a prime example of an *ex ante* influenced behavioral situation is that of ticket scalping. Individuals make a deliberate attempt to enter a queue for economic gain, having

taken into consideration the cost of waiting (Perry & Zarsky, 2013). In addition, this temporal advantage is often a precursor to secondary market activities where the actual monetary value of a resource is substantially higher. In this case, their temporal advantage is used to resell a resource for financial gain. A temporal advantage sometimes does not automatically result in a FIFO selection for service. Medical facilities, particularly emergency services, utilize a triage scheme to assess and care for more critical patients. Perry and Zarsky (2013) also explore several alternatives to FIFO which are characterized the price a party is willing to pay, the need or anticipated enjoyment, individual skill levels and expected service durations. The authors state that as the price one is willing to pay increases, the more expected value that individual presumes to extract from the allocation. In such cases, this can be equivalent to an auction or selling to the highest bidder. In the case for assessing need or expected enjoyment, a third party, or servicer, must judge based on the assumption that all information and facts are accurate (Perry & Zarsky, 2013). In discussing skill, Perry and Zarsky (2013) contend that assessing the level for FIFO allocation can be tedious for even the most virtuous servicers. The final alternative to a FIFO policy maintains that service is offered in an inverse proportion to the expected service time. This application is readily observed in supermarket checkout stands where express lanes are provided for service to customers with fewer items (Perry & Zarsky, 2013). In this case, shorter service times are preferred as an aggregate because of the shortened average wait times and average service times will be reduced. Intrinsic advantages of FIFO offered by the authors suggest FIFO is a scheme where transaction and administrative costs are low (Perry & Zarsky, 2013). This notion is supported by the fact that in some instances, queue occupants themselves enforced the FIFO policy (Brady, 2002). The authors continue their legal discourse by offering four situations where FIFO would be the better option as an allocation scheme (Perry & Zarsky,

2013). First, they suggest a FIFO scheme would be appropriate when resource allocation requires a rapid decision. The primary example given is vehicle traffic at a four-way stop intersection. Allocation time, or delays before proceeding through each stop sign, would be significantly increased if negotiations were involved or comparisons made based on certain relevant criteria (Perry & Zarsky, 2013). Since each vehicle should normally proceed based on arrival time, the first to stop would be the first to proceed. The second suggested situation where FIFO would be appropriate are ongoing or continuous allocations. Perry and Zarsky (2013) give as an example the allocation scheme for a parking lot. Vehicles arrive and fill available spaces until the parking lot is full. The temporal advantage for some results in the reward of finding an available parking spot. Others who either were not able, or chose not to arrive earlier, may forego the opportunity for the resource. If FIFO were not the preferred allocation method in this example, arriving motorists would be delayed until the resources were distributed via another scheme. This could cause unnecessary delay with the potential to create congestion at the entry point until parking spaces were assigned (Perry & Zarsky, 2013). The third setting the authors offer that would be appropriate for a FIFO allocation scheme is situations where all parties would agree to use such a distribution system. Perry and Zarsky (2013) call upon property law to illustrate this point where all parties would likely agree to a FIFO scheme. In this example, securing interest in real property protects both the initial claim and subsequent lien positions in that all creditors would choose to limit their exposure when a debtor offers the same collateral to multiple entities. Future creditors would exercise due diligence in investigating previous claims and either choose to not give credit or enter an agreement willingly while being compensated for the increased risk in the form of higher interest rates (Perry & Zarsky, 2013). Perhaps the last situation the authors suggest as an appropriate FIFO allocation scheme has the most relevance to the current research.

They suggest that a FIFO allocation method would be fitting in those situations where social norms would reject any other mechanism for resource distribution (Perry & Zarsky, 2013). Voluntary cooperation by individuals in a queue can be attributed to social norms and internal enforcement mechanisms, particularly when the queue is visible and well defined (Brady, 2002; Perry & Zarsky, 2013; Zhou & Soman, 2003). Even in the presence of a well-defined queue, exceptions to FIFO can be socially acceptable, such as priority service for certain advantaged individuals and express checkout lanes in the supermarket. The authors suggest that objection to this FIFO exception is even less in a virtual queue where “social norms are almost muted” (Perry & Zarsky, 2013, p. 38). Previous call center research describe queues as *phantoms* in that occupants are invisible to each other as well as to the agents providing service and may well provide a basis for this statement (Brown et al., 2005; Gans et al., 2003; Koole & Mandelbaum, 2002). Even Andrews (2013) recognizes this fact in stating that “On the phone, you are blind” (Andrews, 2013, p. 86). The final portion of Perry and Zarsky's (2013) legal discourse focuses on cases where FIFO can be evaded within social norms. Permitted violations to a FIFO policy can be mutual consent, special need, special merit or special skill (Perry & Zarsky, 2013). Evasion of the FIFO policy can occur when none of the stated exceptions are present and an economic advantage exists for some. The authors state that evasion of a queue can be observed in two forms. The first is where those with an economic advantage can pay for special treatment. The second evasion form is when the same individuals who possess an economic advantage pay others to wait in a queue and obtain resources on their behalf. Socially acceptable exceptions to FIFO do exist but wealth-based allowances can also be present which present the possibility of conflict and has been the subject of much research on queueing theory (Alexander et al., 2012; Brady, 2002; Gurumurthi & Benjaafar, 2004; Maister, 2005; Perry & Zarsky, 2013).

A discussion about queues and fairness would not be complete without a detailed discourse of Rawls' (1971) seminal work titled *A Theory of Justice*. Social justice within queues is predominately associated with a FIFO policy and an associated FCFS system (Alexander et al., 2012; Allon & Hanany, 2012; Larson, 1987). Strict adherence to that policy is usually mandated or breakdown of the system can occur (Larson, 1987; Perry & Zarsky, 2013). Rawls' interpretation of justice is from a perspective that describes social justice as being *justice as fairness* and is closely aligned with social contract theory (Rawls, 1971, p. 16). Rawls' assertion, described through a thought experiment he calls the "original position" (Rawls, 1971, p. 17), is that of a rational group of individuals who were able to decide what form of society they would create, while also being unaware of individual characteristics, social standing or pedigree, would choose a society where rules would be impartial and resources would be equally distributed (Rawls, 1971, p. 19). Rawls states that individuals in the original position are "mutually disinterested" (Rawls, 1971, p. 13) in that they do not pursue selfish interests but are genuinely not interested in each other's pursuits or motivated to benefit each other in any substantial way. This notion put forth by Rawls might aptly apply to customers in a call queue where each person is oblivious to their own position among others waiting for service. Assuming every caller expects to be served in the order of their arrival according to a FIFO policy, the only inequality in this regard would be waiting times of each individual being different due to the exponential distribution nature of arrivals and caller service times. Rawls states there are five formal constraints associated with the concept of what is right that parties in the original position must account for and agree to: (a) all principles should be general, (b) principles should be universal in application, (c) a public conception of justice, (d) impose an ordering on conflicting claims and (e) a finality for adjudication and redress (Rawls, 1971). The principle of generality emphasizes

the absence of any proper names or "definite descriptions" (Rawls, 1971, p. 131). The second constraint imposes a universal application to all and everyone exhibits a willingness to comply. The third condition among the constraints for what is right is the concept of publicity. This supposes the condition for what is right is known to all and is recognized as the arrangement for social cooperation among the population bound by their collective agreement. Ordering on conflicting claims is a constraint that implies there is a definite priority for conflicting claims and the ordering should be "transitive" (Rawls, 1971, p. 134) in precedence so not to appeal to "force and cunning" (Rawls, 1971, p. 134). The last constraint enforces a binding recognition of the principles as the last "court of appeals" where "no higher standards to which arguments in support of claims can be addressed" (Rawls, 1971, p. 135). Rawls' theories rely on what he describes as a rational person. His assumptions on what constitutes a rational person is someone who is not affected by envy, regard life in its entirety and are concerned for their future prospects beyond any temporal boundaries (Rawls, 1971). Rawls asserts individuals in the original position are also under the influence of what he describes is the "veil of ignorance" (Rawls, 1971, p. 136). This condition assures individuals do not know their particular social standing, pedigree or level of education when deciding as a community on the type of society they wish to form. Individuals in a call queue exhibit a level of blindness as to who or how many others occupy the same realm. Even Perry and Zarsky (2013) invoke Rawls' conceptual thought experiment in stating "Blindness to irrelevant differences is reminiscent of the Rawlsian 'veil of ignorance,' . . ." (Perry & Zarsky, 2013, p. 13). Rawls' assumption is that each person would choose resources fairly and impartially in the original position, constrained by the veil of ignorance and call queues have appropriately been described as *phantom* queues by some researchers due to occupants being unaware of others in the queue (Brown et al., 2005; Gans et al., 2003; Koole &

Mandelbaum, 2002). Callers within a queue are most likely unaware of who or how many customers are ahead of them in the queue, how long it will likely be before they receive and may correctly assume, in most cases, the calls within the queue are answered in a FCFS manner through a FIFO policy (Larson, 1987). Although individuals within a call queue can never realistically develop a collective original position agreement likened to a Rawlsian call queue society, they are restricted to the order imposed upon them by the servicer and accept the obligatory nature of waiting alone for service, since it likely fulfills a fundamental need and will provide some form of future value. His theory on justice and his foundational thought experiment on original position and veil of ignorance present some intriguing views on the applicability in describing the fundamental nature of a call queue. Rawls develops his theory further in stating there are two principles of justice that would be selected by individuals in the original position. These two principles are that “each person is to have an equal right to the most extensive liberty compatible with a similar liberty for others” and “social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone’s advantage, and (b) attached to positions and offices open to all” (Rawls, 1971, p. 60). Rawls’ first principle is egalitarian in nature and supposes everyone has an equal right to the fundamental liberties enjoyed by all. A call queue with a FIFO policy can exhibit characteristics of Rawls’ first principle. Customers who call a service center may be presented with information regarding their expected wait. After the wait time announcement, a customer may choose to remain in the queue for service or balk at the announcement (Gans et al., 2003; Jouini et al., 2011). Regardless of whether wait time information is provided to each customer or not, those who remain in the queue enjoy the equal liberty of the prospect for service at some point in time in the future. In a FIFO call queue universe, the singular fundamental liberty of service might be

thought of as a manifestation of Rawls' first principle of justice. Rawls' second principle extends his framework of social justice by defining appropriate wealth distribution among those in the original position and behind the veil of ignorance. Rawls develops the first part of the second principle by stating "The intuitive idea is that the social order is not to establish and secure the more attractive prospects of those better off unless doing so is to the advantage of those less fortunate" (Rawls, 1971, p. 75). Rawls elaborates further by stating what he calls the *difference principle* which is designed to regulate inequalities to favor those that are worse off. Rawls' difference principle does not seek to remove all inequalities in a social and economic structure, but rather allows those inequalities to exist only if it benefits those in a lesser position. This concept might easily be applied to the FIFO policy for service in a call queue. For example, each customer in a queue arrived at different times and their total wait will vary according to when they arrived. The inequality that exists can be the moment a servicer becomes available in a call center. A FIFO policy will automatically route the caller who has waited the longest to the next servicer available. This strategy allocates an inequality of service for only one person to the next occupant in the queue. The next person waiting in the queue can be considered to be in a lesser position, since they have been waiting longer than anyone else. Although an argument can be made that the next person in the queue, and the one who has waited longer, enjoys an earned equity position for service, a service in random order scheme would violate the previously stated call queue liberty and wealth arguments (Alexander et al., 2012). A FIFO call queue manifests many of the social justice principles put forth in Rawls' thought experiments. Rawls' theory describing justice as fairness, mixed with the social contract aspect of the original position, provide striking theoretical propositions for studying customer call service queues from a social justice perspective.

But what happens when social justice and equity collide? For example, if a servicer can create a priority queue where customers can pay for faster service, the egalitarian nature of a Rawlsian original position FIFO policy may be violated, especially if others in the queue become aware of the perceived injustice. In a call queue where the veil of ignorance exists, invoking a priority queue can be accomplished by providing a separate phone number for priority customers or the servicer resolving a caller's phone number and routing them for faster service. In this situation, those utilizing the priority queue have more information than the rest of the customers and Rawls' social justice idea of justice as fairness is violated. However, those who have paid for priority service expect some reward for their equity contribution. Alexander et al., (2012) investigated priority queues, or multilevel queues, and their effect on consumers. Their approach was to understand and apply social justice and equity theory in a mixed-methods study of theme park ride queues. Their focus was to determine the impact of priority queues on consumers occupying both a normal and a priority queue (Alexander et al., 2012). The authors state that priority queues have received little study even though they are becoming more prevalent within various service industries. In understanding the social justice aspect of their study, Alexander et al., (2012) invoke Rawls' concept of justice as fairness in describing social justice implications for waiting in line. The authors address the equity aspect of their study by citing Glass and Wood's (1996) definition of equity as that being the balance between what is offered and what is received in exchange relationships, much like that of paying for priority service in a queue. In their mixed method approach, the researchers utilized a sequential exploratory survey to build a foundation for their follow-on quantitative segment. The initial study revealed some consumers felt bitterness toward customers in the priority queues and regarded themselves as being a victim of injustice. Others felt there was a benefit to priority queues, especially when they felt the utility

of not waiting in line was worth the investment (Alexander et al., 2012). From the initial exploratory survey, hypotheses were developed to investigate contemporary queueing theory effects of a priority queue on customers in the main queue and value expectations for customers paying for priority service. The results of their first experiment indicated increasing queue lengths have a negative effect on customer satisfaction and loyalty thereby affecting word of mouth purchase recommendations. The second experiment contradicted the findings of the first experiment in that increasing queue lengths had a positive effect on customer satisfaction, word of mouth recommendations and purchase intentions when in the presence of longer queue lengths and priority queues were available (Alexander et al., 2012). Finally, it was found that satisfaction mediates the relationship between the independent variables of waiting time and cost of priority queues and the dependent variables of word of mouth recommendations and purchase intent. Further implications of their research produced a contradiction in that achieving social justice for those in a main queue denies the equity exchange for those in a priority queue. In addition, achieving an equitable exchange for those in a priority queue will produce injustice for those in the main service queue (Alexander et al., 2012).

With regard to social norms in queues, Allon and Hanany (2012) investigated queues through the application of a novel game theory approach which attempts to understand why customers cut in line and why others allow such behavior. Their main assertion is that queues are in effect a social system and therefore within the realm of game theory where rational players are seeking to maximize their individual utility. Their results showed that during a single iteration of the game, an equilibrium is achieved and FIFO emerges as a self-regulating mechanism where cutting in line is rejected. When multiple iterations of the game are played, a priority rule, in the form of $c\mu$ where c is waiting costs and μ is expected service rate, is sustained which allows a

system manager to prioritize service based on the $c\mu$ value which improves overall system performance (Allon & Hanany, 2012). When players can exercise a strategy based on queue length, it is discovered that a priority scheme develops. More urgent requests, and those with very little work content, are allowed to cut in line when the queue length is below a certain threshold (Allon & Hanany, 2012). Finally, their study concluded that welfare maximizing equilibria occur when constant monitoring of the queue is provided and thorough punishment applied for queue deviations (Allon & Hanany, 2012).

Call Center Operations

Traditional call center operations are primarily focused on maintaining a balance between service quality and operational efficiency (Borst et al., 2004; Brown et al., 2005; Koole & Mandelbaum, 2002; Mandelbaum & Zeltyn, 2009). Service quality is critical to building and maintaining strong customer satisfaction and long-term relationships and operational efficiency attempts to minimize costs for a particular level of resource utilization (Aksin et al., 2007; Ibrahim & L'Ecuyer, 2012; Zohar et al., 2002). Most calls are typically inbound but outbound calls are conducted for customer follow-up after service, marketing or collections activities (Holman et al., 2007; Mehrotra & Fama, 2003; Shen, 2010). In many call centers, some agents are trained to handle calls requiring different skill sets and some call centers have specialized agents to conduct outbound calls as well (Ali, 2010).

Call Center Growth

The emergence of the contact center, primarily attributed to associated technological advances such as the internet, are a natural evolution in contact channels for conducting customer service operations (Gans et al., 2003; Koole & Mandelbaum, 2002). Contact centers have rapidly expanded over the last decade and expected growth continues for the future

(Deloitte, 2015). In an effort to reduce costs and seek efficiencies, some businesses have adopted a shared services approach to common business functions (Aksin & Masini, 2008). Outsourcing of important jobs such as call center support are common practice in an attempt to reduce operating costs after carefully considering in-house alternatives (Aksin & Masini, 2008; Holman et al., 2007). The costs for maintaining a customer support facility are very high and human resources account for a large portion of call center expense and some researchers maintain the cost of call center human resources consume anywhere from 60% - 80% of the operating budget (Aksin et al., 2007; Bapat & Pruitte, 1998; Brown et al., 2005; Shen, 2010). Over the last decade on average, call centers have experienced an annual 20% growth (Aksin et al., 2007; Bouzada, 2009; Fukunaga et al., 2002; Jouini, Dallery, & Aksin, 2009; Koole & Mandelbaum, 2002; Mandelbaum & Zeltyn, 2009). The International Customer Management Institute, an industry association comprising insiders, trainers and consultants, indicates the voice channel accounts for approximately 65% of contact center interactions and predict other contact channels, such as social media, mobile and web chat, may one day surpass voice as a preferred means of customer contact. Call centers are ubiquitous and represent the primary means of customer contact for many businesses (Bouzada, 2009; Brown et al., 2005; Demiriz et al., 2009; Weinberg et al., 2007). Holman, Batt, and Holtgrewe (2007), in their collaborative scholar account titled *The Global Call Center Report: International Perspectives on Management and Employment*, provided survey data from almost 2,500 call centers from 17 countries that employed over 475,000 individuals.

Call Forecasting

Call forecasting is another critical component of call center operations (Akhtar & Latif, 2010; Ibrahim & L'Ecuyer, 2012; Koole & Mandelbaum, 2002; Shen, 2010; Weinberg et al.,

2007). Calls arriving into a call center exhibit a stochastic process but statistics for the arrivals, wait times and service times can be determined over finite time intervals (Koole & Mandelbaum, 2002). Given the probability that callers may call at any particular minute within a finite interval, call arrivals exhibit a Poisson distribution (Koole & Mandelbaum, 2002). The most fundamental operational call center model is the M/M/1 queue, which assumes a Poisson arrival process with an exponentially distributed service time and a single server (Borst et al., 2004; Koole & Mandelbaum, 2002; Shen, 2010; Zohar et al., 2002). From a service quality perspective, call centers primarily follow an 80/20 rule where 80% of the calls are answered within 20 seconds (Ali, 2010; Koole & Mandelbaum, 2002; Mathew & Nambiar, 2013; Whitt, 2005). Customer impatience is a primary factor in call reneging and abandonment that can influence call volume forecasts if not taken into consideration (Aguir et al., 2004; Akhtar & Latif, 2010; Gans et al., 2003). In comparing Erlang models, only the Erlang-A model accounts for customer abandonments even though the Erlang-C model is predominantly used for call center forecasting (Angus, 2001; Borst et al., 2004). Technologies to improve customer experience when faced with a queue include providing an IVR option to complete simple transactions and announcing estimated call wait times before entering the queue (Aksin et al., 2007; Jouini et al., 2011; Koole & Mandelbaum, 2002; Mathew & Nambiar, 2013). Customers can also request a callback if the announced wait time exceeds their limit on how long they want to wait (Aksin et al., 2007; Gans et al., 2003).

Personnel Planning

Call center operations management addresses both short and long-term resource acquisition and deployment (Aksin et al., 2007). Resource scheduling is made several weeks or even months in advance and updates are made based on forecasts for inter-day and intra-day

intervals (Shen, 2010). Future workload for any planning period must be appropriately matched to the resources needed to satisfy the demand. This workload is referred to as *offered load* and is defined as the product of the call arrival rate and the mean service time of the arrivals (Shen, 2010). Aksin et al. (2007) assert call centers face resource deployment challenges aimed at addressing staffing, scheduling and routing problems. Call centers collect operational data to serve in creating forecasts of call volumes over short intervals, typically 15, 30 and 60 minutes (Aksin et al., 2007). Call center resources are scheduled to minimize cost while achieving a certain service level and customer satisfaction, which may include some customers having to wait in a queue (Aksin et al., 2007; Atlason et al., 2004; Fukunaga et al., 2002). The staffing problem deals with how many *servers*, or CSRs, are required to achieve the desired level of service over a given time interval within three modes of operation: quality driven, efficiency driven and a rationalized system (Borst et al., 2004). Borst et al.(2004) also investigated the applicability of the square-root safety staffing principle for M/M/s queues that derives optimum staffing levels based a function of the ratio of customer waiting costs and service agent costs. The Erlang-C model does not allow call blocking and therefore if the staffing levels were not adequate to handle calls as they arrive, a service queue will develop. Some research has been conducted on customers who balk before a queue, or abandon once in the service queue and subsequently retry their call, and that can lead to distortions in forecasting and staffing calculations (Aguir et al., 2004; Mandelbaum & Zeltyn, 2007). When call volume is unpredictable and there are more arrivals than forecast, a queue will develop where callers will have to wait and their perceived wait time may exceed their expectation (Maister, 2005). Depending on cost, several operational models are utilized which are *quality-driven* and *quality and efficiency-driven* regimes (Robbins et al., 2010). In order to adjust to unanticipated call

volumes, or surges in expected arrivals, a *square-root safety-staffing rule* is implemented within a quality and efficiency-drive regime and is indicated by $s = a + \beta\sqrt{a}$ where a is the offered load and β represents the desired service grade (Gans et al., 2003). In addition to the staffing problem, call center managers face a scheduling problem in determining the optimum number of shifts to be worked to satisfy service level objectives or labor requirements (Aksin et al., 2007; Ali, 2010; Fukunaga et al., 2002; Ibrahim & L'Ecuyer, 2012; Koole & Pot, 2006). Determining the right agent to satisfy a particular customer's needs is regarded as the call routing problem (Aksin et al., 2007). This particular challenge arises within a multi-skill call center with constraints around which agent must service a particular call. Atlason et al. (2004) use simulation and a novel cutting plane algorithm to minimize staffing costs within a multi-skill call center environment. The staffing and scheduling problem becomes more complex due to the time-inhomogeneous Poisson process. Many researchers have investigated this phenomenon in an attempt to forecast call volumes (Brown et al., 2005; Gans et al., 2003; Robbins et al., 2010).

Performance Measures

Calculation of most performance measures for call centers assumes a system that has reached a steady state (Garnett et al., 2002; Whitt, 2005). Aksin et al. (2007) recognize that a portion of the call volume may be retrials to a system where the caller abandoned the queue or perhaps not received a resolution on the first attempt. First-call resolution and speed of service are considered key performance indicators for call centers and CSRs. (Aksin et al., 2007; Deloitte, 2015; Koole & Mandelbaum, 2002; Koole & Pot, 2006).

Call center demand modulation may become necessary because of highly unpredictable and time-varying volume of calls (Akhtar & Latif, 2010; Aksin et al., 2007; Ibrahim & L'Ecuyer, 2012). Modulating call center demand is referred to as *load balancing* and it can provide various

means to deal with unpredictably high call volumes (Aksin et al., 2007; Gans et al., 2003; Koole & Mandelbaum, 2002) . One of the simplest mechanisms for load balancing is call admission (Aksin et al., 2007). When this operation is imposed, a caller will receive a busy signal and will not be allowed to enter the call queue. This outmoded treatment of callers can lead to poor customer satisfaction and they will likely attempt a retrial in the future (Akhtar & Latif, 2010; Aksin et al., 2007; Shen, 2010).

Advances in technology and telecommunications have enabled call queues to become more interactive and offer a better degree of perceived and actual customer service. Many businesses employ callback queues where customers can choose not to wait in the queue and be called back when a customer support representative becomes available (Aksin et al., 2007; Gans et al., 2003).

Queue Simulations

Simulations have been widely used in analyzing and understanding complex systems within the realm of operations research (Winston, 1994). Call queues are one area where much investigation has been focused on using simulation where some argue Erlang formulas and queuing theory fail to provide a complete understanding of queue dynamics (Akhtar & Latif, 2010; Ali, 2010; D. R. Anderson et al., 2013; Bapat & Pruitte, 1998; Bouzada, 2009; De Mesquita & Hernandez, 2006; Mathew & Nambiar, 2013; Mehrotra & Fama, 2003). The predominate approach to queue simulation is through a Discrete Event Simulation (DES) approach (Akhtar & Latif, 2010; Bertoli, Casale, & Serazzi, 2009; De Mesquita & Hernandez, 2006; Leong, 2007; Mathew & Nambiar, 2013). Mesquita and Hernandez (2006) describe a unique approach to DES using spreadsheets and Visual Basic for Applications (VBA) code. Their demonstration offers a unique method for using inherent spreadsheet functionality for data

analysis and software code to manage the actual simulation. Simulations are also considered valid experimental designs in that they are specifically created to represent the natural environment where processes and activities normally occur (Sekaran & Bougie, 2013).

Exploratory Research

Social justice implications within queues has been frequently studied and human behavior within that context has produced varying degrees of observable phenomenon (Andrews, 2013; Larson, 1987; Maister, 2005). However, very little investigation has occurred regarding telephone *phantom* queues in relation to perceived or observable social justice constructs. In order to understand perceived justice within queues, an exploratory study is appropriate to gain understanding of the phenomenon under question (Sekaran & Bougie, 2013).

A novel approach to survey deployment has emerged over the last decade that has allowed researchers to reach a more international audience through technology advances such as the internet and crowdsourcing platforms (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). MTurk is an online marketplace where workers can select and complete tasks for a relatively small amount of compensation. Each task is referred to as a Human Intelligence Task (HIT) and compensation depends on nature of the task and desired qualifications of the worker. The use of MTurk for research is gaining momentum with some researchers finding practical advantages such as reduced costs with relatively good quality data (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). The composition of MTurk workers has been demonstrated to not vary significantly from other survey respondents (Huff & Tingley, 2015).

Research Methodology

Overall Approach and Rationale

The two most generally practiced and accepted research methodologies are quantitative and qualitative. Both methods are appropriate for conducting business research and carefully chosen based on the nature of the problem and previous research found in a review of the literature (Cooper & Schindler, 2011; Creswell, 2012; Sekaran & Bougie, 2013). Quantitative research is normally guided by a need to explain relationships, understand why an event occurs, analyze trends in a series of data or attempt a precise measure of something in order to predict future outcomes (Cooper & Schindler, 2011; Creswell, 2012). Creswell (2014) also emphasizes the importance of a thorough literature review for a quantitative approach. A comprehensive literature review allows the researcher to justify the research effort as something worthy of investigation by stating a definitive need for understanding the problem in greater detail or even discover a solution (Cooper & Schindler, 2011; Creswell, 2012).

In business research, quantitative methods prevail in investigating a variety of concerns for companies. A quantitative approach is appropriate for research within the finance, accounting and management disciplines (Creswell, 2012). The nature of business research seeks to provide answers to guide leaders in making informed decisions about their challenges and seek better ways to bring value to the customers (Sekaran & Bougie, 2013). Given that much research within business is suitable for a quantitative approach, the foundation for a quantitative design is based on the scientific approach to investigation (Sekaran & Bougie, 2013). The fundamental principles of scientific research include purposiveness, rigor, testability, replicability, precision, objectivity, generalizability and parsimony (Sekaran & Bougie, 2013). This research study took great measure in ensuring all the principles were addressed, but particularly the principles of

purposiveness, rigor, precision, and parsimony. The methodology chosen was determined to be appropriate to satisfy all principles and address the research questions as stated. A comparison between selected elements of quantitative and qualitative methodology is made in Table 1

Table 1

Quantitative Versus Qualitative Research

	Quantitative	Qualitative
Research Focus	Describe, explain, or predict	Understand, interpret
Researcher Involvement	Limited, controlled	Researcher is a participant
Research Purpose	Describe, predict, and test theory	Understanding, theory building approach
Sample Size	Large	Small
Research Design	Determined beforehand, single, or mixed-method approach	Can evolve, multiple methods
Data Analysis	Statistical and mathematical	Human analysis and quantitative coding, contextual phenomenon

Creswell (2014b) emphasizes that not only does a researcher choose between quantitative, qualitative or a mixed methods approach, they must also decide "on a type of study within these three choices (Creswell, 2014b, p. 12). Quantitative research typically relies on survey or experimental designs and qualitative research is segmented into narrative or phenomenological designs (Creswell, 2014). An experimental approach can be either a laboratory experiment or a field experiment where the phenomenon under investigation can be observed in the natural environment (Sekaran & Bougie, 2013). Another form of quantitative investigation that has emerged within business research is simulation (Cooper & Schindler, 2011; Sekaran & Bougie, 2013). Simulations are valuable in that they allow the creation of an environment that represents

reality and allows investigation of a real-world system as it evolves over time (Cooper & Schindler, 2011; Sekaran & Bougie, 2013; Winston, 1994).

Research strategy. The design for this study centered on understanding the effectiveness of a VPQ implementation on reducing the wait time for select individuals in a service queue. Prior to conducting the simulation experiment, an investigation on the impact of long service queue wait times and perceived justice within a queue were investigated as antecedents to service queue abandonment. An exploratory survey instrument was developed to assess the relationship between customer patience, perceived justice, and intent to abandon a service queue.

Hypotheses. Hypotheses were developed to appropriately address the stated research questions. Considering the proposed VPQ is designed to reduce the wait time for a select population determined by a servicer, extended wait times and their impact on an individual's patience threshold as an antecedent to service queue abandonment was investigated. In addition, since service queue abandonment for apparent queues has been investigated from a social justice perspective (Allon & Hanany, 2012; Brady, 2002; Larson, 1987; Perry & Zarsky, 2013), the role of perceived queue justice as an antecedent to queue abandonment was also studied. Although a VPQ implementation would be resigned to only call queues, the first two hypotheses were researched for both apparent and phantom service queues (line-standing and call queues). Therefore, the following hypotheses were proposed to investigate customer intentions on abandoning a service queue:

H_1: Patience Threshold (PT) is negatively related to Intent to Abandon (IA) a service queue.

H_2: Perceived Justice (PJ) is negatively related to Intent to Abandon (IA) a service queue.

The following hypothesis was proposed for the simulation experiment to investigate VPQ efficacy:

The call wait times for the VPQ will be significantly less than the normal service queue.

H_3: $W_{VPQ} < W_{SQ}$

H_{3a}: $W_{VPQ} \geq W_{SQ}$

Methodology

Sekaran and Bougie (2013) substantiate the applicability of simulation by stating, “Simulation uses a model-building technique to determine the effects of changes, and computer-based simulations are becoming popular in business research” (Sekaran & Bougie, 2013, p. 188). Creswell (2014) straightforwardly states, “An experimental design is the traditional approach to conducting quantitative research” (Creswell, 2014, p. 294). Therefore, a DES was employed utilizing a personal computer with a spreadsheet software package with multiple simulation runs to assess wait times for those customers in the VPQ and those in the normal service queue. A next-event time-advance mechanism was selected as the simulation protocol which allows the simulation to advance after each imminent event which in this case, was the next arrival into the queue (Winston, 1994).

Population Selection

Setting. The quantitative survey portion of this research was accomplished online using Amazon’s MTurk platform. Respondents chose whether to complete the survey and were compensated for their effort. The computer simulation portion of the VPQ was conducted on a personal computer using a standard spreadsheet software application. Ten simulation runs were accomplished and assessed for VPQ effectiveness. Simulation results were captured individually for statistical analysis of normal service queue and VPQ wait times.

Participants. Participants for the exploratory survey instrument were volunteer Human Intelligence Task workers within the Amazon Mechanical Turk platform within the United States. Participants were compensated for their time in completing the survey and appropriate controls were placed on the quality and reliability of MTurk respondents with only those workers with a HIT reliability rating of 95% or higher being allowed to participate.

Research Instruments

Survey instrument development followed a deliberate process modeled on the design phases presented by Cooper and Schindler (2011) shown in Figure 12.

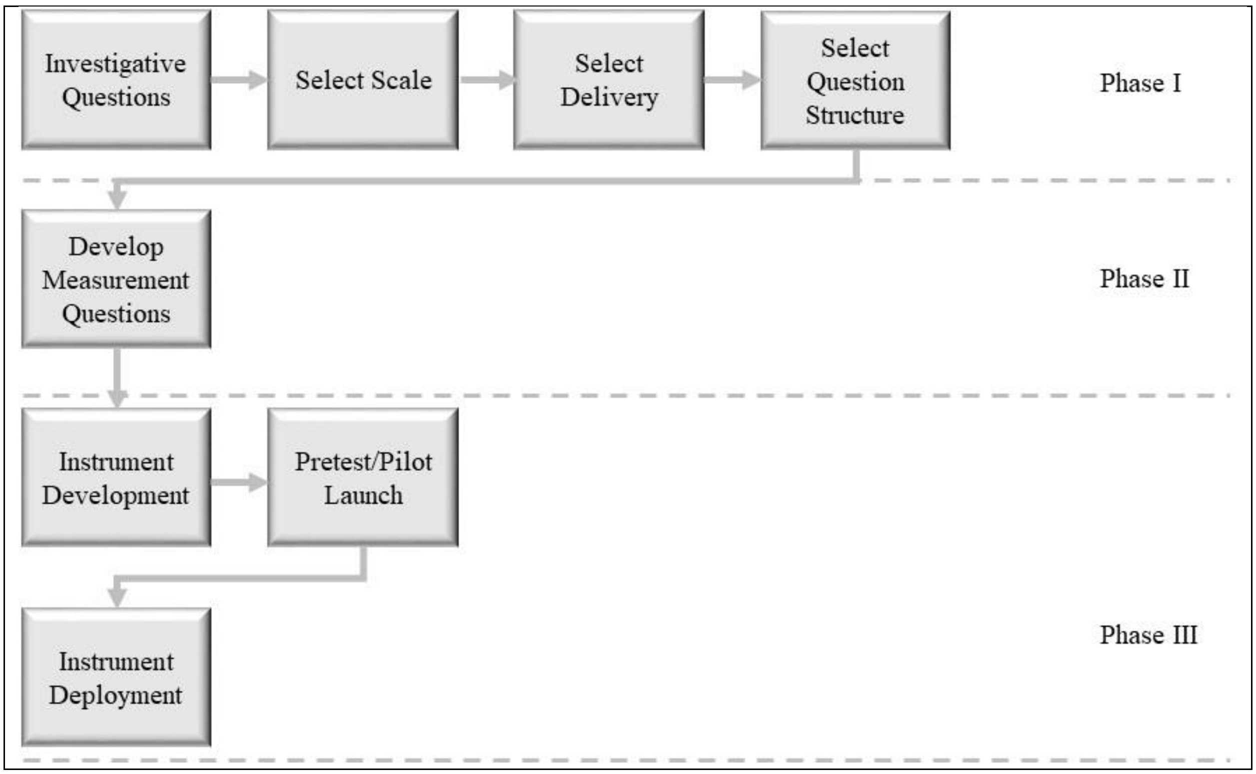


Figure 12. Deliberate process model of Cooper and Schindler (2011).

This approach allowed a seamless transition from the initial investigative proposal to the final survey instrument. Scale selection, delivery method and question structure occurred in the initial phase of instrument development. The second phase allowed for detailed development of the items associated with the constructs being studied. The final phase included a pilot launch of the

survey instrument followed by modification for deployment. The exploratory study survey instrument was launched via the MTurk platform and respondents were compensated for their participation in the study. Survey deployment to only those MTurk workers who resided in the United States was done to focus on that demographic for perceptions related to the research questions on social justice constructs within service queues.

The survey instrument was designed to include two parts that measure perceived justice and patience thresholds within both line standing and call service queues. The survey contained a scenario description for each type of service queue for respondents to consider when answering the questions and the questions were randomized in order to minimize common method variance (Eichhorn, 2014).

Each of the 10 DES simulations were conducted with 100 events, or customers entering the queue. Invoking the VPQ operation was deterministic with every fifth entry being an advanced reservation position where a simulated arrival would be placed at a later time. VPQ entries were filled by averaging the wait times for all previous entries into the normal service queue and filling the advanced reservation that many time periods into the future after creation.

Protection of Human Subjects: Ethical Considerations.

A statement of informed consent will be provided to MTurk workers who respond to the survey in accordance with the University of the Incarnate Word policy and 45 CFR §46.116. The University of the Incarnate Word Institutional Review Board approved the survey associated with this study (#17-07-001) and the researcher was trained in human subjects research and obtained Institutional Review Board approval shown at Appendix A. A notice was presented to each respondent before beginning the survey with language that indicated the estimated time to complete the survey was no more than 15 minutes. Participants received assurances that their

data would confidential and the survey data did not include names and other personally identifiable information. The survey did not ask subjects questions of a sensitive nature and there was no physical risk or expense related in participating. Notification was also given that respondents were free to discontinue the survey at any time.

Data Collection

Data for studying the relationship between patience thresholds, perceived justice, and intent to abandon a service queue was collected via an online survey instrument hosted on a cloud-based Google Drive platform utilizing Google Forms for the actual instrument. Survey data were collected automatically and the end results were exported as a common spreadsheet file extension for analysis. The data were saved to a secure cloud storage location for further examination. Computer simulation data were collected after each of the 10 simulation runs and saved to a spreadsheet application on a cloud platform for further analysis of the different wait times.

Data Analysis

Data collection results. A total of 299 responses were gathered from the survey deployment on the MTurk platform. The data were then analyzed for outliers and incomplete responses. A total of 17 responses addressing perceptions of call queues contained missing data and 18 responses on perceptions of waiting in a line contained missing data. These responses were considered ignorable data and were determined to be missing completely at random (MACR) and therefore deleted listwise (Hair, Black, Babin, & Anderson, 2010; Peugh & Enders, 2004). The final dataset was composed of 281 valid responses on perceptions of call queues and 280 valid responses on perceptions for waiting in a line. The data were then analyzed for descriptive statistics and reverse coded items were recoded into the appropriate values.

Data analysis for the 10 computer simulation runs was conducted utilizing a statistical analysis software package to determine significant differences in hypothesized call wait times between two call queue populations.

Findings

Perceptions of Service Queues – Exploratory Survey

Data cleaning. A pilot survey was launched through Amazon's MTurk platform to assess the items developed in an initial research model on perceptions on service queues. The pilot survey addressed both line standing and call service queues. The initial theoretical model is shown in Figure 13.

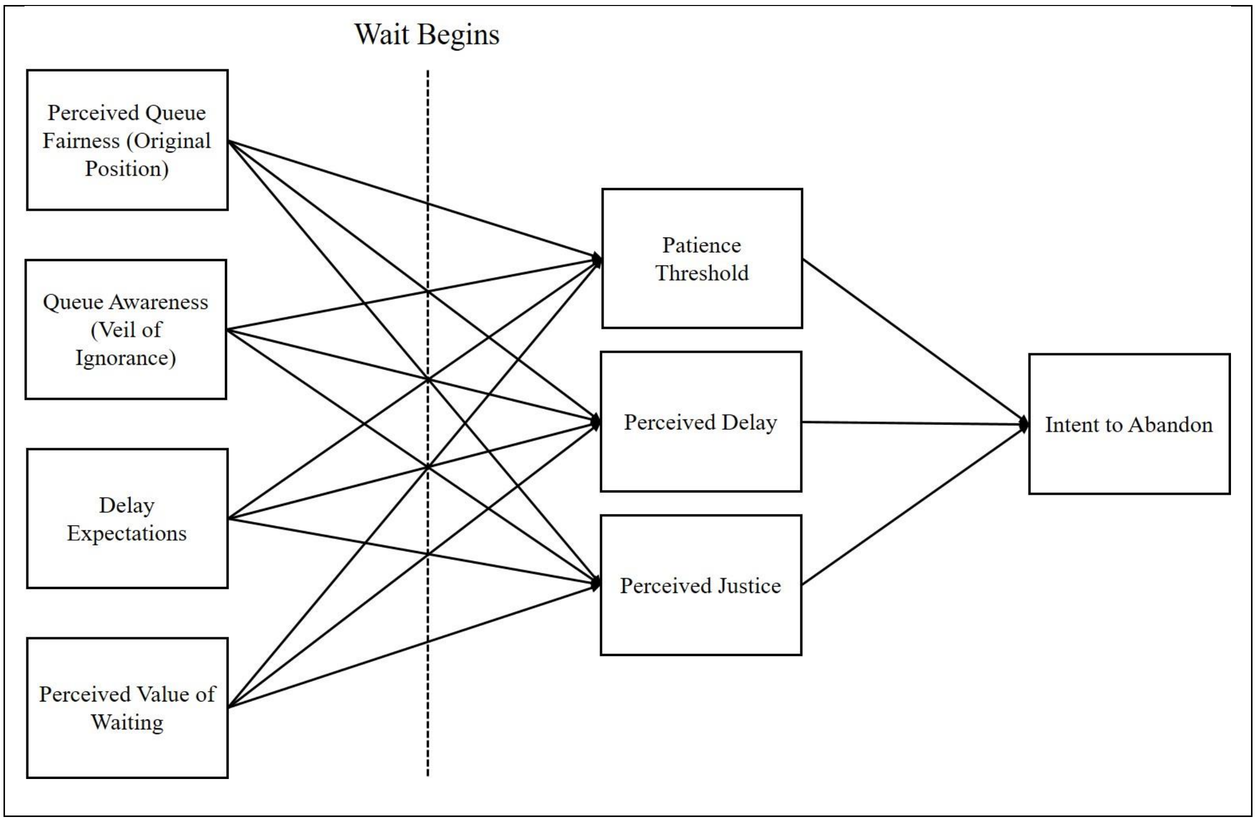


Figure 13. Initial theoretical model (Pugh, 2017).

Data analysis. The pilot launch collected a total of 55 responses and after data analysis and listwise deletion of missing and outlier (straight-lining) data, 48 valid responses for perceptions of line standing queues and 50 valid responses for perceptions of call queues remained. The data were then analyzed using a common statistical analysis software package to recode the reverse coded items. A list of coded items for each construct is shown in Table 2.

Table 2

Coded Construct Items

Construct	Code
Perceived Queue Fairness (Original Position)	QF
Queue Awareness (Veil of Ignorance)	QA
Delay Expectations	DE
Perceived Value of Waiting	PV
Patience Threshold	PT
Perceived Delay	PD
Perceived Justice	PJ
Intent to Abandon	IA

Results analysis. Evaluation of the constructs within SmartPLS (Ringle, Wende, & Will, 2005) indicated several item loadings associated with QA, DE, PV and PD were very low and were removed from the actual survey instrument. A deliberate effort was made to focus on the research questions at hand and adjust the theoretical model accordingly. After further revision, the theoretical model was refined as depicted in Figure 14.

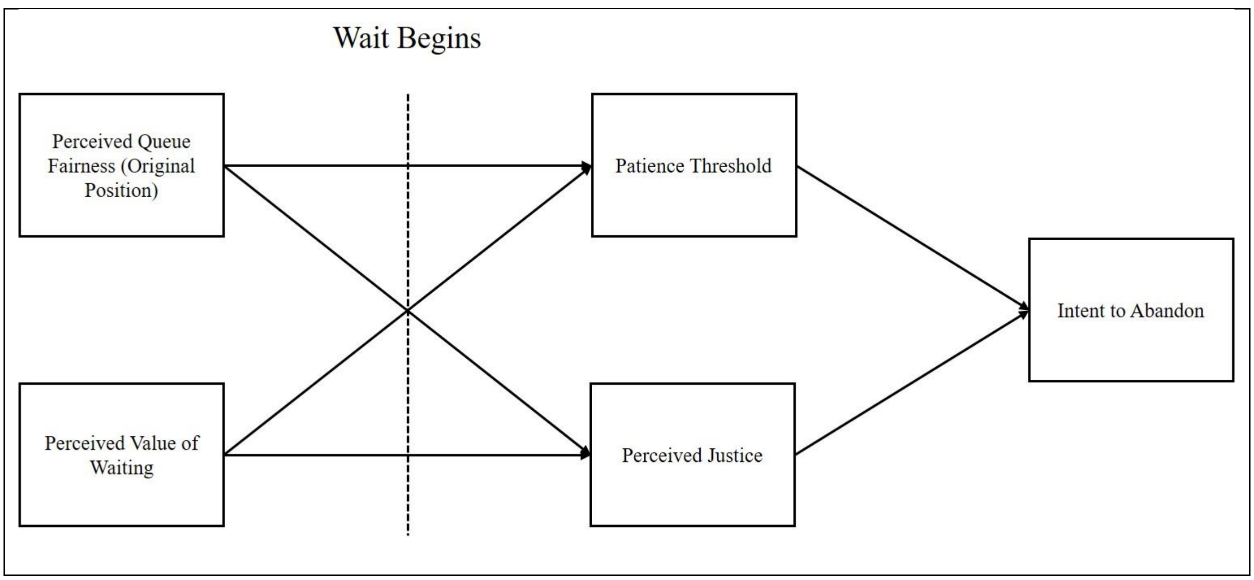


Figure 14. Final theoretical model (Pugh, 2017).

Although the final survey instrument included items for QF and PV, as depicted in the measurement model, the resultant research model only contained constructs aligned with the

research questions on PT, PJ, and IA, shown in Figure 15. The survey instrument used is included in Appendix B. Data on the excluded constructs will be used for future analysis and continued model validation.

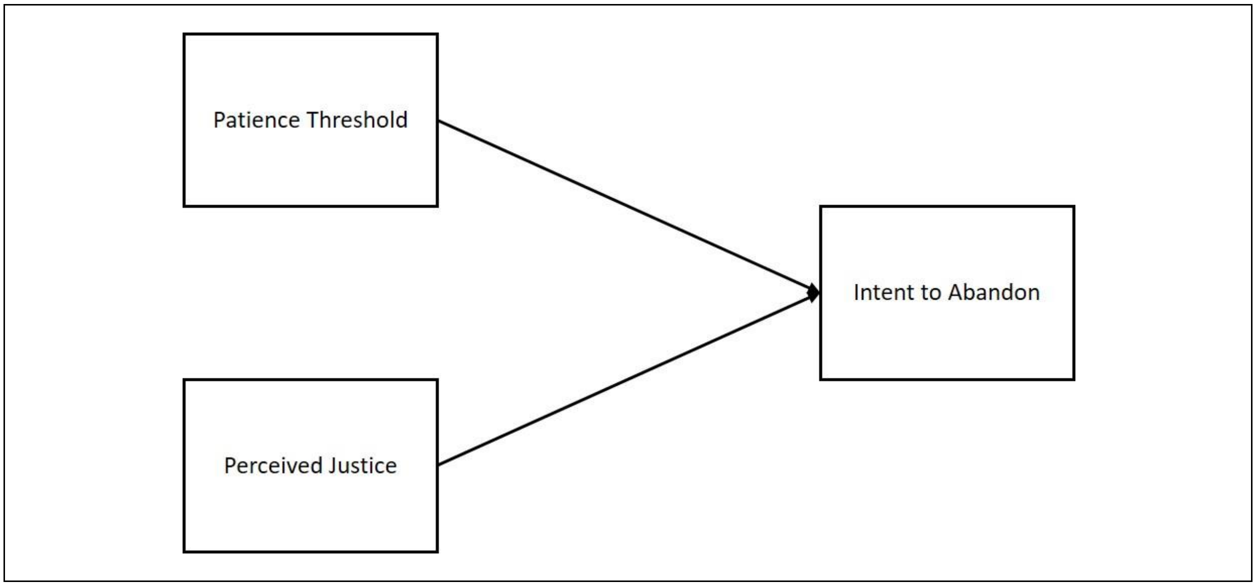


Figure 15. Research model (Pugh, 2017).

Perceptions of Service Queues – Line Standing

Data cleaning. 299 responses were gathered from the survey deployment on the MTurk platform. The data were then analyzed for outliers and missing values and 18 responses on perceptions of waiting in a line contained missing data. These were determined to be ignorable data and were deleted (Hair, Black, Babin, & Anderson, 2010; Peugh & Enders, 2004). The final dataset contained 280 valid responses on perceptions for standing in a service line. The data were then analyzed for descriptive statistics and reverse coded items were recoded into the appropriate values.

Results analysis. A two-step process was used to evaluate the measurement model and the resultant structural model proposed in this exploratory research (Anderson & Gerbing, 1988). The first step determines the reliability and validity of the constructs and the second step

evaluates the structural model for predictive power and magnitude of relationships (Henseler, Ringle, & Sinkovics, 2009). Evaluation of the proposed model included both discriminant and convergent reliability. Convergent validity is an indicator that related constructs converge, or have a high degree of variance in common (Hair, Black, Babin, & Anderson, 2010). All items in each construct for the final structural model indicated a high degree of correlation, which is an indicator of good convergent validity (Templeton & Burney, 2017). Discriminant validity is a measure of each construct being distinct from other constructs (Hair et al., 2010). All factor loadings were greater than .70, which is an indicator of good discriminant validity (Henseler et al., 2009). Additional discriminant validity is provided by the Fornell-Larcker (1981) criteria being satisfied, which is the square root of the Average Variance Extracted (AVE) being greater than that item with all other latent variables, as shown on the diagonal of Table 3. Internal consistency is indicated by composite reliability being higher than .70 for each factor (Hair, Ringle, & Sarstedt, 2011). Cronbach's Alphas were generally good with the exception of PT, which was .620.

Table 3

Line Standing Item Loadings, AVEs, CR and Factor Correlations

Item loadings			AVEs, CR, factor correlations					
Factor	Item	Loadin g	AVE	CR	Cronbach's Alpha	PT	PJ	IA
Patience Threshold (PT)	PT1	.784	.568	.798	.620	.754		
	PT2	.714						
	PT5	.762						
Perceived Justice (PJ)	PJ2	.949	.765	.866	.719	.025	.875	
	PJ5	.793						
Intent to Abandon (IA)	IA2	.916	.845	.916	.817	.350	.231	.919
	IA3	.923						

Although the commonly acceptable threshold is .70, values from .60 to .70 are at the "lower limit of acceptability"(Hair et al., 2010). All AVEs were greater than .50 which is considered good for convergent validity (Hair et al., 2011; Hair et al., 2010). The Variance Inflation Factor (VIF) for all constructs was less than 5 which indicates multicollinearity is not a factor (Hair et al., 2011). The R^2 value was .17 and the final structural model is shown in Figure 16.

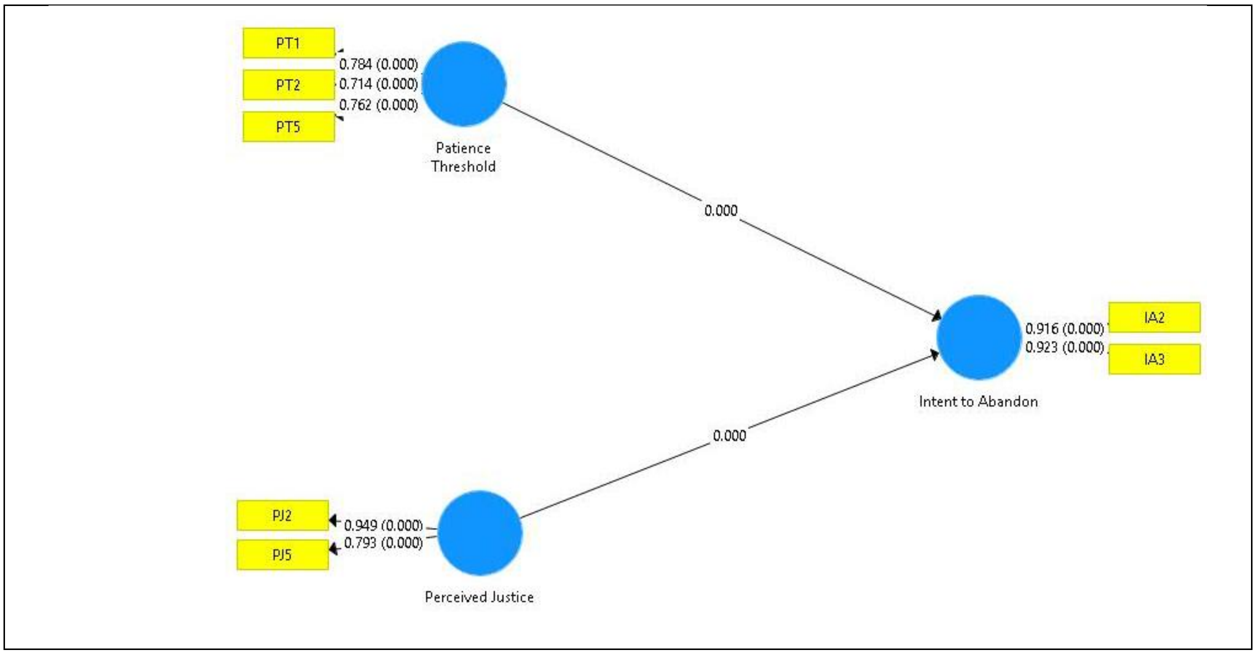


Figure 16. Line standing structural model (Pugh, 2017).

PJ and IA retain only two items each, but the strength of the relationship indicates dependable results and Yong and Pearce (2013) consider a factor with two items reliable if they are highly correlated (> .70).

Perceptions of Service Queues – Call Queues

Data cleaning. 299 responses were gathered from the survey deployment on the Mturk platform. The data were then analyzed for outliers and missing values and 17 responses on perceptions of waiting in a line contained missing data and again these were determined to be ignorable data and were deleted (Hair, Black, Babin, & Anderson, 2010; Peugh & Enders, 2015).

The final dataset contained 281 valid responses on perceptions for waiting in a call queue. The data were then analyzed for descriptive statistics and reverse coded items were recoded into the appropriate values.

Results analysis. As with the data for perceptions on standing in a line, the factor loadings for perceptions of call queues were greater than .70. The Fornell-Larcker (1981) criteria was satisfied as shown in Table 4. Composite reliability was higher than 0.70 for each factor (Hair et al., 2011). Cronbach's Alpha for PT was again lower than .70, but meets the lower limits of reliability (Hair et al., 2010).

Table 4

Call queue item loadings, AVEs, CR and factor correlations

Item loadings			AVEs, CR, factor correlations					
Factor	Item	Loading	AVE	CR	Cronbach's Alpha	PT	PJ	IA
Patience Threshold (PT)	PT1	.730	.602	.819	.682	.776		
	PT2	.763						
	PT5	.830						
Perceived Justice (PJ)	PJ1	.766	.667	.909	.875	.163	.817	
	PJ2	.796						
	PJ3	.830						
	PJ4	.833						
	PJ5	.856						
Intent to Abandon (IA)	IA2	.925	.84	.913	.810	.396	.239	.917
	IA3	.908						

All AVEs were greater than .50 and the VIF for all constructs was less than 5. The R^2 value was .19 and the final structural model is shown in Figure 17. Only IA again contained two items, however the factors were highly correlated (> .70).

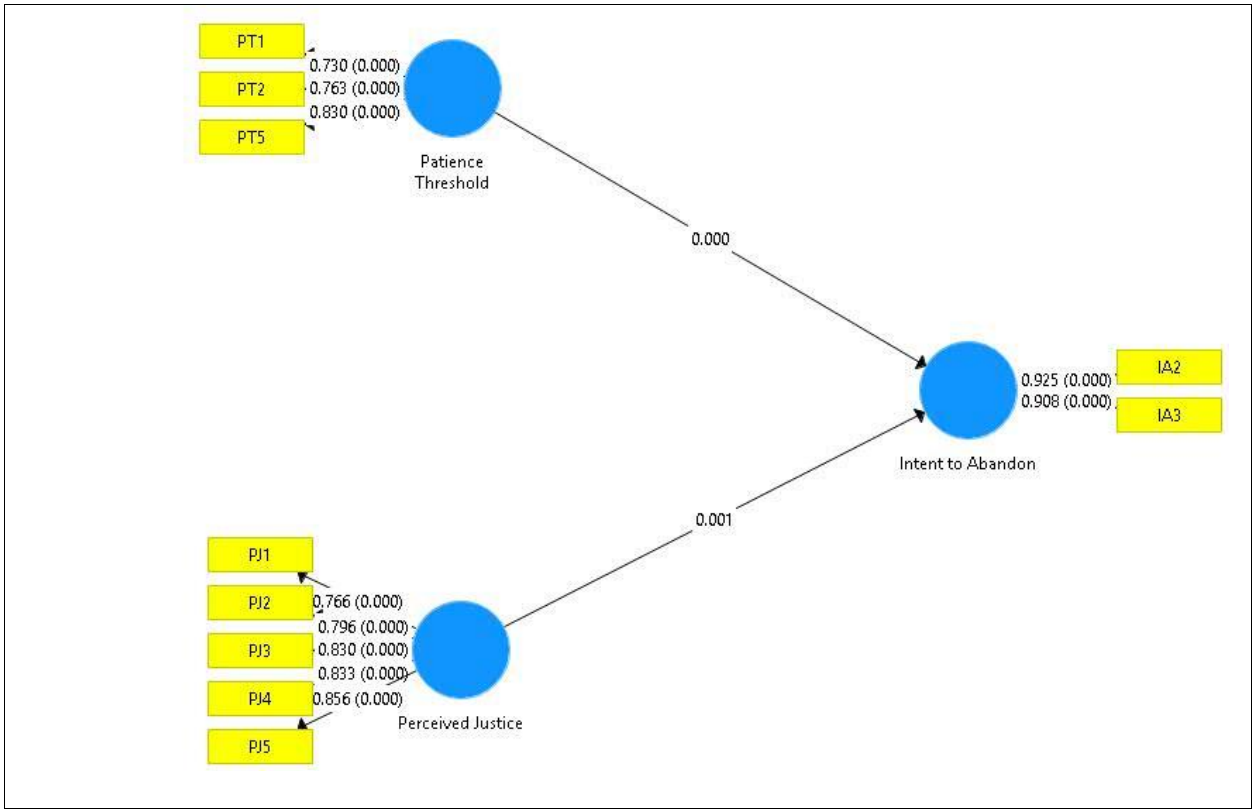
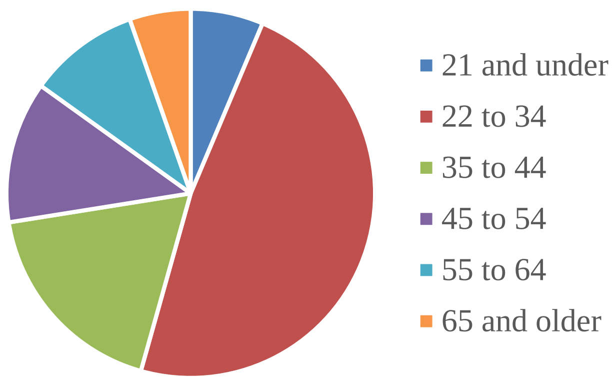


Figure 17. Call queue structural model (Pugh, 2017).

Demographics. Respondent age ranges are depicted in Figure 18. The majority of the respondents were in the 22 to 34 age range consistent in part with the findings of Ipeirotis (2010) with 64% being male and 36% female. The education level of the respondents is indicated in Figure 19. A substantial proportion (37.7%) reported having a bachelor's degree. This finding is consistent with that of Paolacci et al. (2010) who attribute the higher levels of education to early adopters of new technology such as Amazon's MTurk platform.

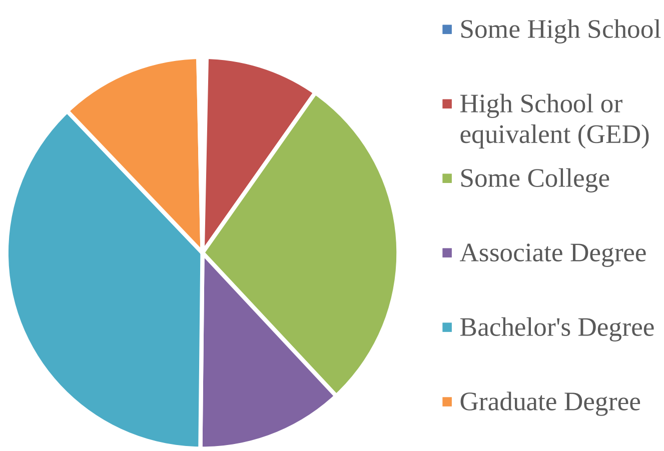
Respondent Ages



Age Range	Count	Percent
21 and under	19	6.38%
22 to 34	143	47.99%
35 to 44	54	18.12%
45 to 54	37	12.42%
55 to 64	29	9.73%
65 and older	16	5.37%

Figure 18. Respondent ages.

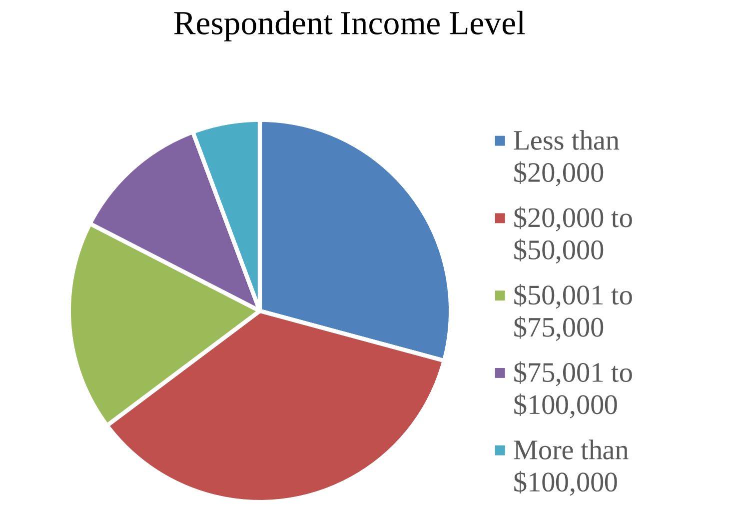
Respondent Education Level



Education Level	Count	Percent
Some High School	1	.34%
High School or equivalent (GED)	28	9.43%
Some College	84	28.28%
Associate Degree	36	12.12%
Bachelor's Degree	112	37.71%
Graduate Degree	35	11.78%
Doctoral Degree	1	.34%

Figure 19. Respondent education level.

The reported income levels at Figure 20 are also consistent with the findings of Paolacci et al. (2010) in finding that the although the education level were higher than the general population, the income levels reported were lower than the average.



Income Level	Count	Percent
Less than $20,000	87	29.19%
$20,000 to $50,000	106	35.57%
$50,001 to $75,000	53	17.79%
$75,001 to $100,000	35	11.74%
More than $100,000	17	5.70%

Figure 20. Respondent income level.

Discrete Event Simulation

Results analysis. For the first DES run, the wait times for the normal service queue were not normally distributed. Normality is a fundamental assumption for many multivariate analysis tests and must be present in order to use the *t* statistic to evaluate differences in sample means (Hair et al., 2010). A two-step method was used to transform the data in order to achieve normality (Templeton & Burney, 2017). The first step in the transformation required a fractional rank order of the cases be created and the second step utilized a normal inverse distribution function to achieve normality (Templeton & Burney, 2017). The Kolmogorov-Smirnov test for normality values for the initial dataset and the transformed values are shown in Figure 21

	Kolmogorov-Smirnov		
	Statistic	df	Sig.
Wait Time (Normal Queue)	.228	18	.014*
Wait Time (Normal Queue Transformed)	.125	18	.200

Figure 21. Kolmogorov-Smirnov values before and after data transformation.

The first DES run is depicted in Figure 22 and showed a significant difference in wait times between customers who would have occupied a VPQ (M = 14.65, SD = 9.61) and customers in the normal service queue after the data were transformed (M = 32.85, SD = 15.91); $t(34) = 4.15$, $p = .001$. This simulation run had two advanced reservations that went unfilled and the filled to total reservations ratio was .90 or 10% loss of available VPQ positions.

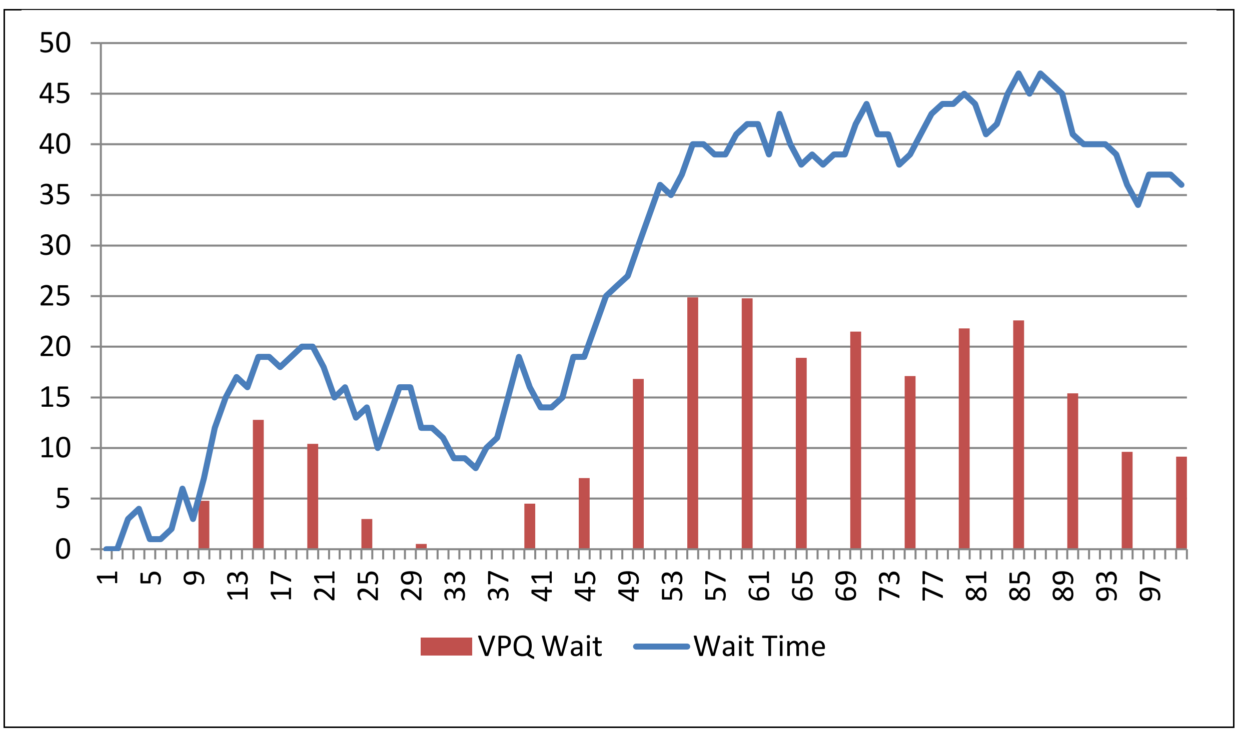


Figure 22. First DES simulation run.

The behavior of this simulated single server queue showed an initial rise in wait time and then a gradual decrease before the wait time rose again until the simulation was terminated at the one hundredth arrival. The wait time and service time of the first 20 customers is depicted in Figure 23. There is a brief moment where the server is idle for four time periods between when the first customer finishes service and the second customer arrives.

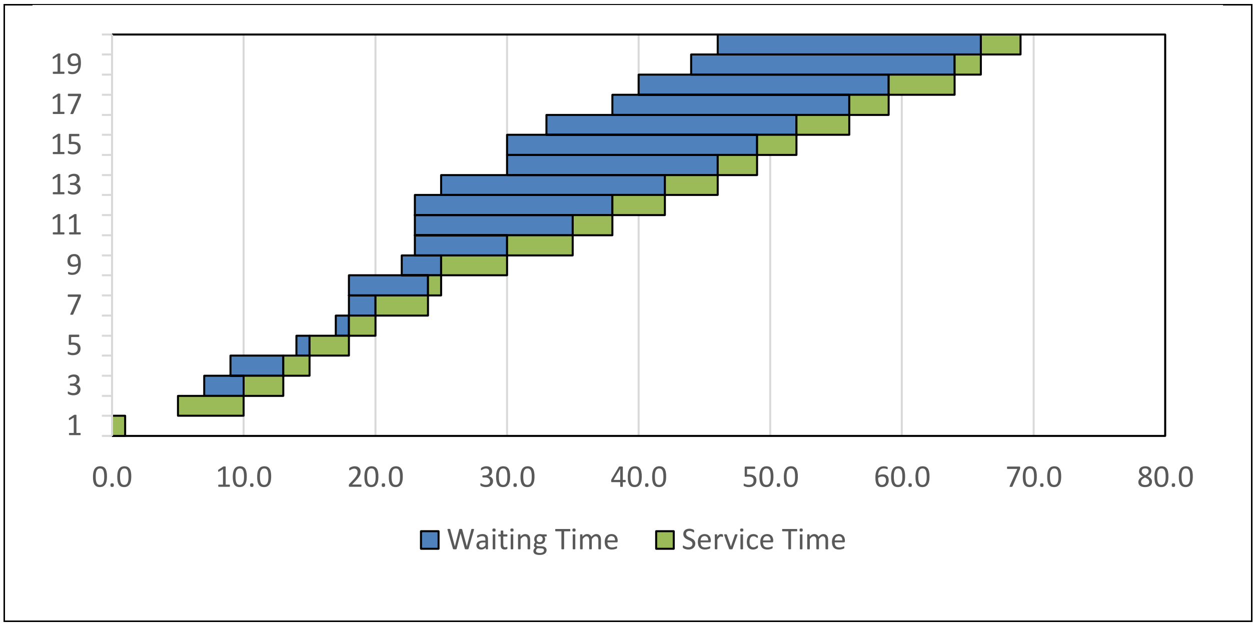


Figure 23. First DES simulation run first 20 customers.

The second DES run shown in Figure 24 indicated a significant difference in wait times between customers who would have occupied a VPQ (M = 16.46, SD = 6.84) and customers in the normal service queue (M = 39.80, SD = 16.71); $t(38) = 5.78$, $p = .001$. This simulation run is characterized by a rapidly increasing wait time after the fifteenth customer and the wait time steadily increases until the simulation is terminated after one hundred arrivals. It is noted that when wait times for the service queue steadily increase, the difference between wait time of the VPQ and the service queue increase over time as well. This is due to the programmed parameters of the simulation averaging the wait times previous to the advanced reservation and populating the advanced reservation with a customer at the creation time of the advanced reservation plus

the average wait time. This iteration of a VPQ simulation had no loss of the 20 available advanced reservations.

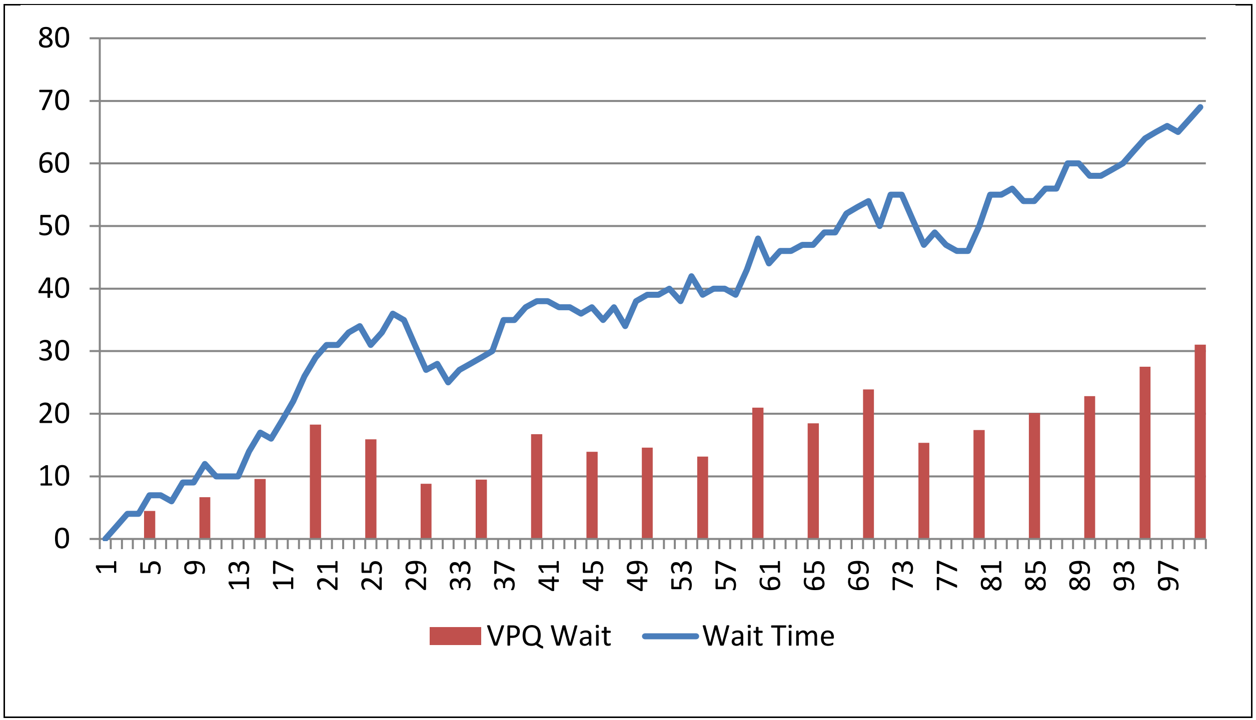


Figure 24. Second DES simulation run.

The rise in wait time for the first 20 customers can be seen in the bar graph in Figure 25.

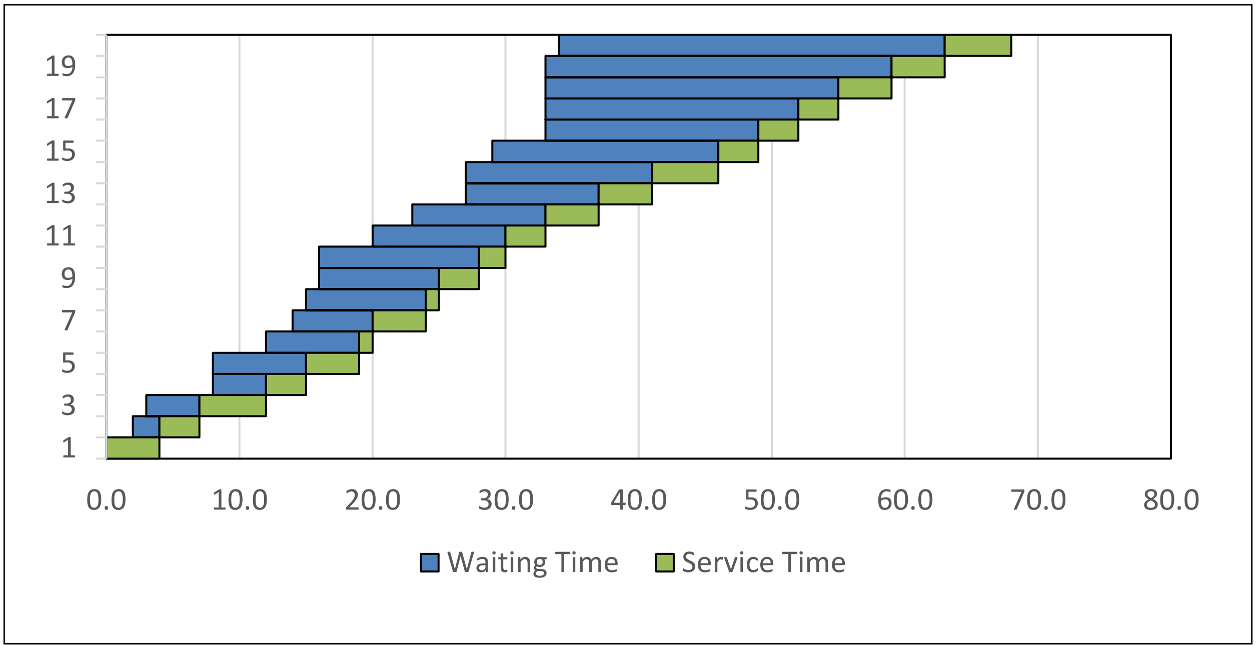


Figure 25. Second DES simulation run first 20 customers.

The third DES run also indicated a significant difference between wait times for customers who would have occupied a VPQ (M = 18.39, SD = 10.52) and customers in the normal service queue (M = 31.63, SD = 16.19); $t(30) = 2.74$, $p = .010$, but at a slightly lower level as shown in Figure 26. This simulation run had a loss of four available advanced reservations with a filled to total reservations ratio of .80 or a loss of 20%. This simulation run exhibited an initial rise in wait time and then a decrease to zero wait time for arrivals between 26 and 33. Due to the initial random decrease in service times and longer arrival periods, the remaining advanced reservations showed a smaller differential between wait times and VPQ wait times. This was due to the VPQ being filled based on average wait time up to the creation of the follow-on advanced reservation and the lower wait time average impacted the efficiency of the VPQ for decreasing wait times.

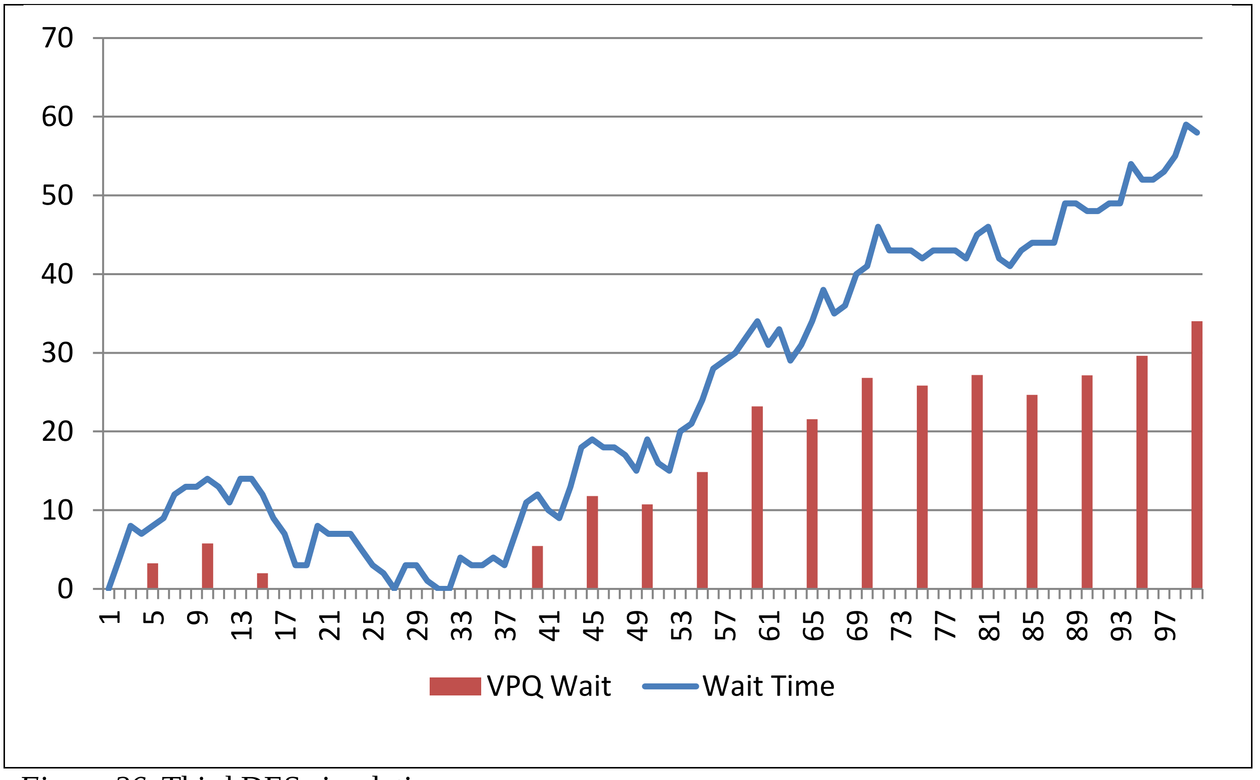


Figure 26. Third DES simulation run.

The bar graph of the first 20 arrivals is shown in Figure 27 and indicates the initial rise in wait times leading up to the decrease to zero at several points before approaching 60 time units at the end of the simulation.

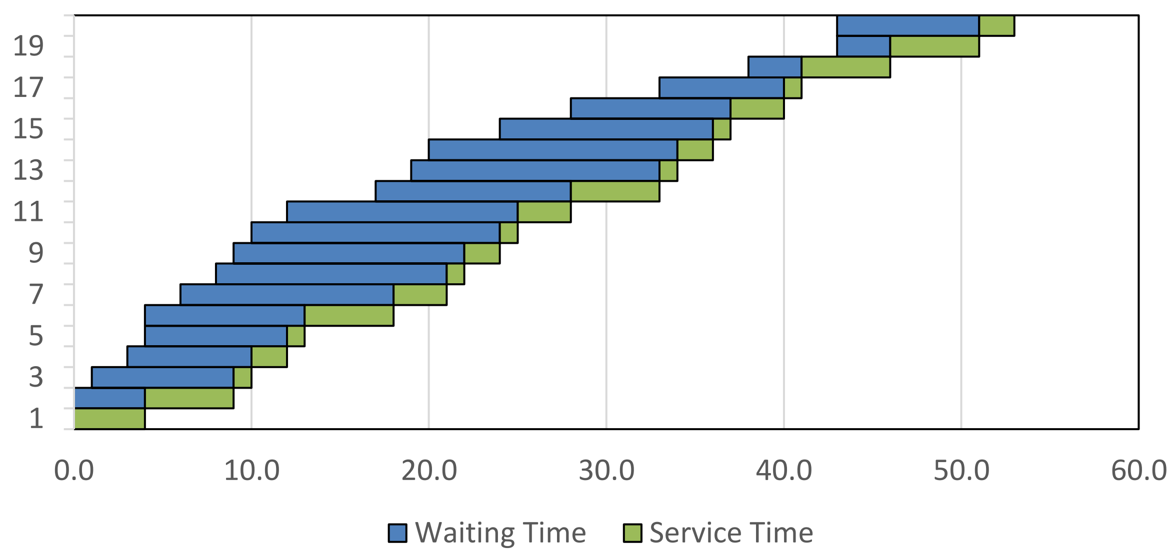


Figure 27. Third DES simulation run first 20 customers.

The fourth DES run at Figure 28 showed there was a significant difference between customers who would have occupied a VPQ (M = 7.16, SD = 4.53) and customers in the normal service queue (M = 25.75, SD = 11.10); $t(30) = 6.66$, $p = .001$. This simulation run is an appropriate example of extremely varying wait times across the complete time domain. As a result of the erratic wait time, the averaging the previous wait times to determine when to populate the advanced reservation of the VPQ caused a loss of one reservation at the beginning of the simulation, two halfway through the run and another at the 80th arrival. The filled to total reservations ratio was again .80 or a loss of 20% of available advanced reservations.

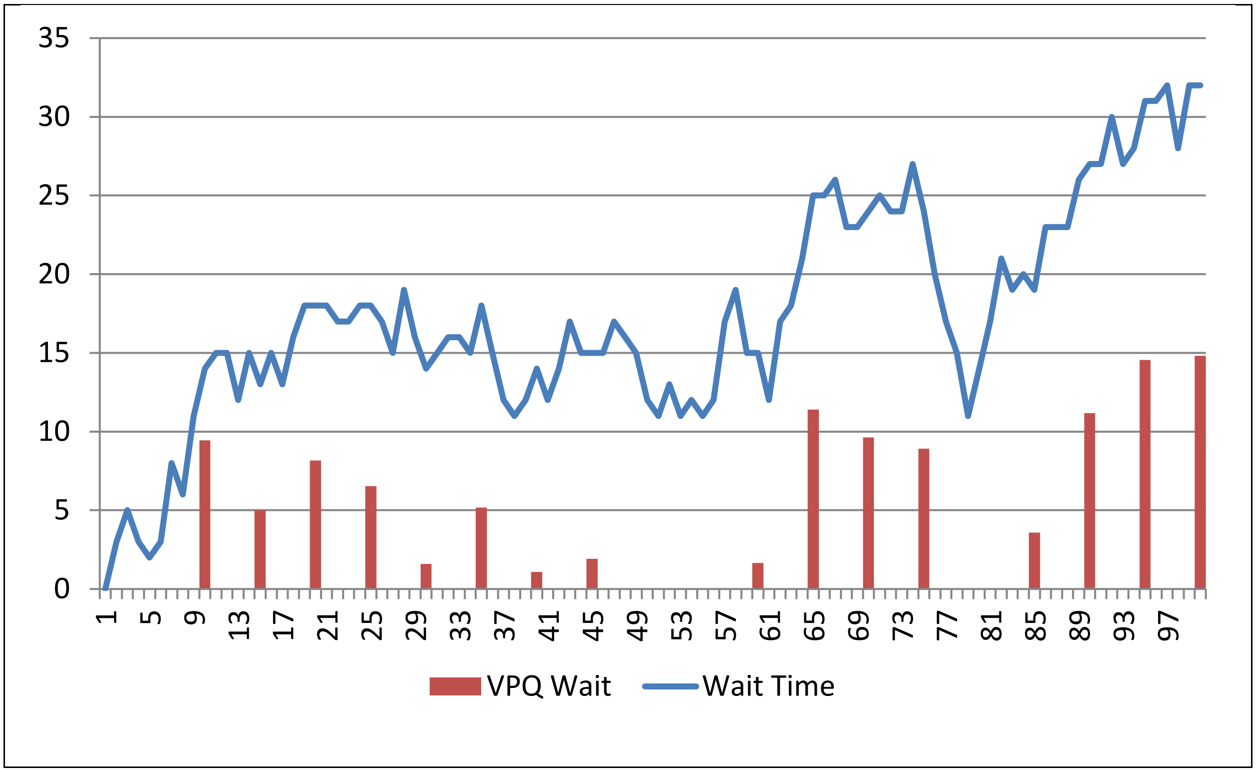


Figure 28. Fourth DES simulation run.

The bar graph of the first 20 arrivals at Figure 29 indicates a normal rise in wait times and a short wait period for arrival number 5 where the first advanced reservation went unfilled.

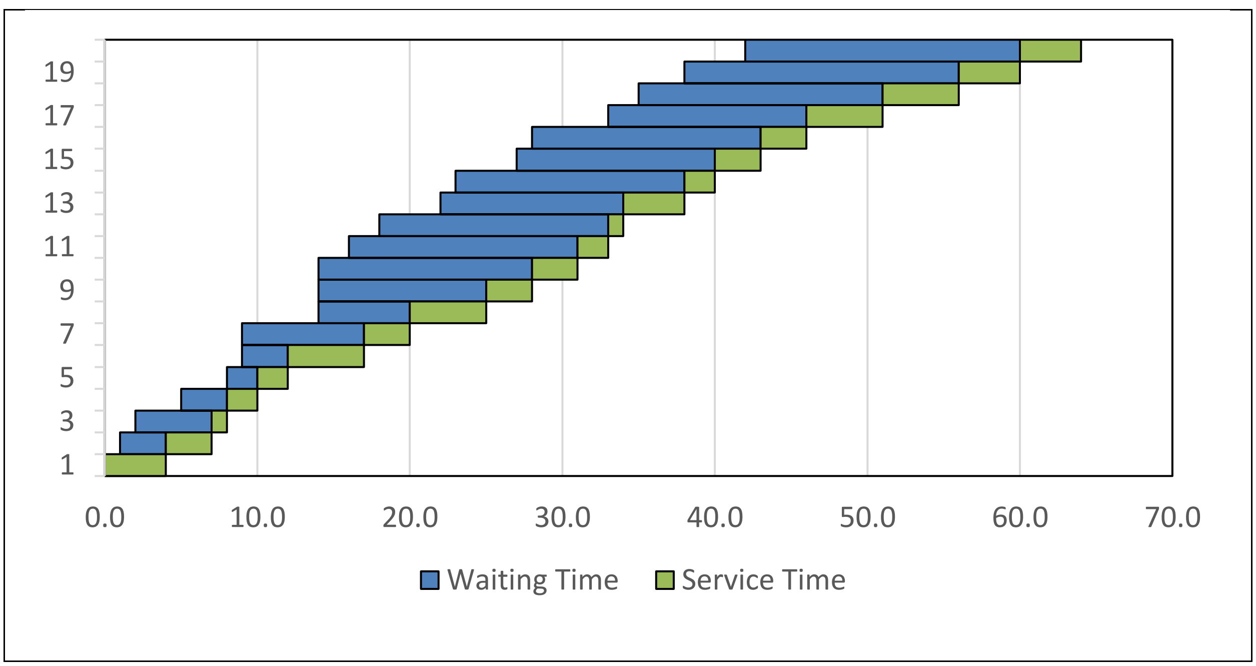


Figure 29. Fourth DES simulation run first 20 customers.

The fifth DES run shown in Figure 30 indicated a significant difference between customers who would have occupied a VPQ (M = 9.95, SD = 6.04) and customers in the normal service queue (M = 20.06, SD = 6.28); $t(38) = 5.59$, $p = .001$. The beginning of this run saw no wait for the first fours arrivals and due to the average being zero, the first advanced reservation created in the fifth arrival position had the same wait time as the same position in the normal queue. The filled to total reservations ratio was 1.00 and there were no advanced reservations that went unfilled.

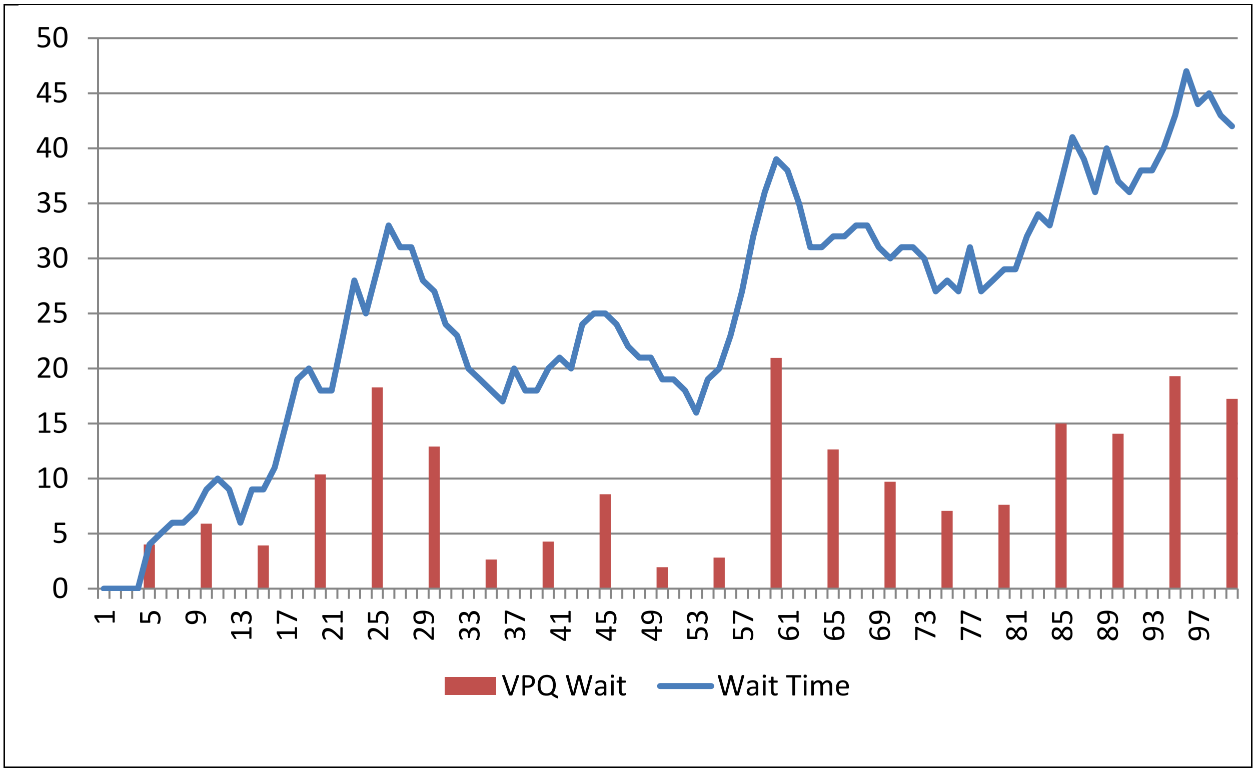


Figure 30. Fifth DES simulation run.

Another interesting behavior of this particular simulation run was that in addition to the first four arrivals having zero wait time, positions four through seven arrived at the same time. This contributed to an immediate increase in wait time due to a single server operation. The bar graph at Figure 31 for the first 20 arrivals indicates the initial zero wait time and four simultaneous arrivals.

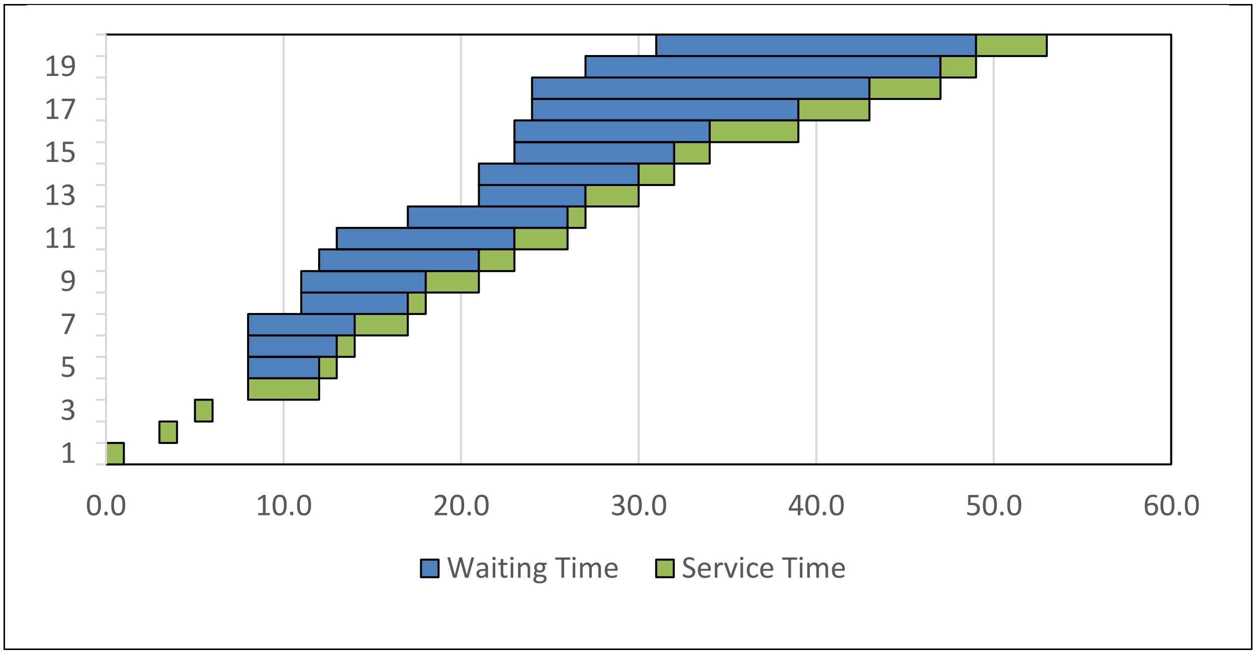


Figure 31. Fifth DES simulation run first 20 customers.

The sixth DES run shown in Figure 32 also indicated a significant difference between customers who would have occupied a VPQ (M = 22.64, SD = 13.74) and customers in the normal service queue (M = 35.82, SD = 22.24); $t(32) = 2.08$, $p = .046$ but at a lower level than the third run. This is attributed to most arrivals before the 25th customer seeing no wait time for service. The filled to total reservations ratio was .85 or a loss of 15% of advanced reservations. The bar graph of the first 20 arrivals for run number six, shown in Figure 33, indicates many early arrivals have no wait time. Even with a loss of 15% of the available advanced reservations, the difference in wait time between the normal queue and advanced reservation was still significant, $p < .05$.

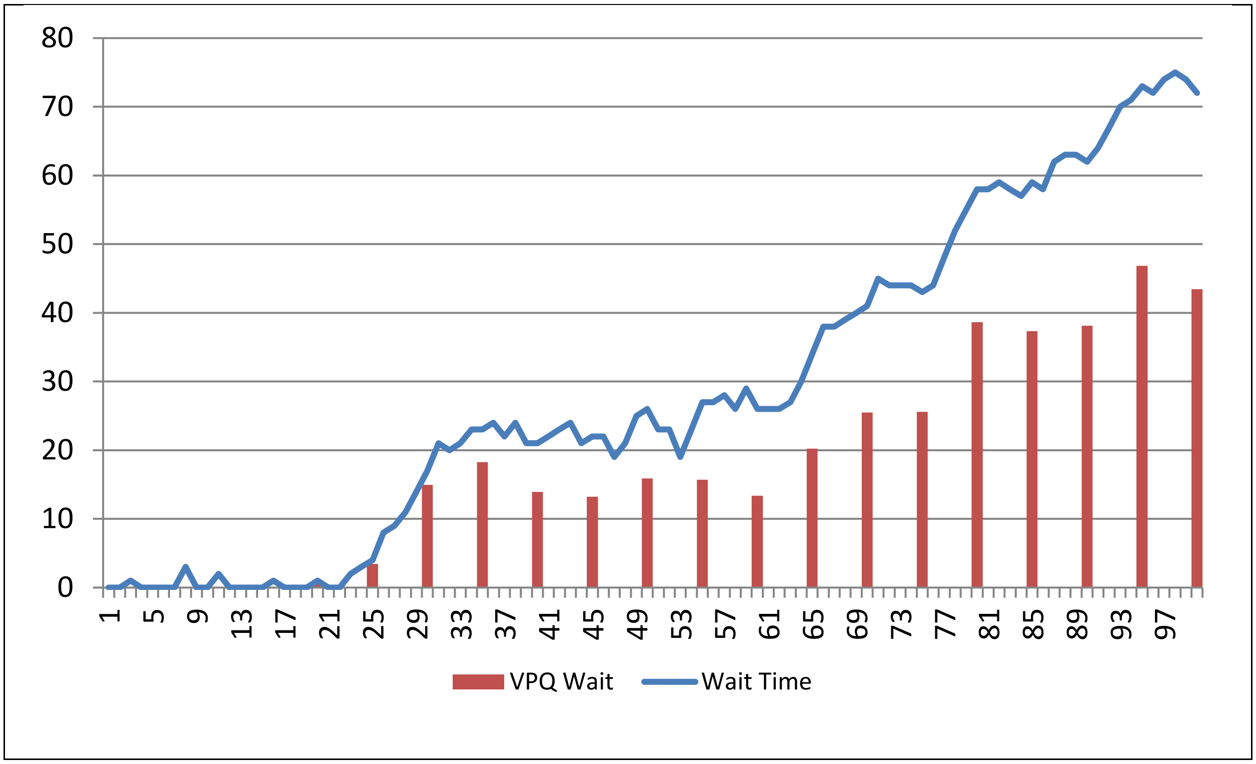


Figure 32. Sixth DES simulation run.

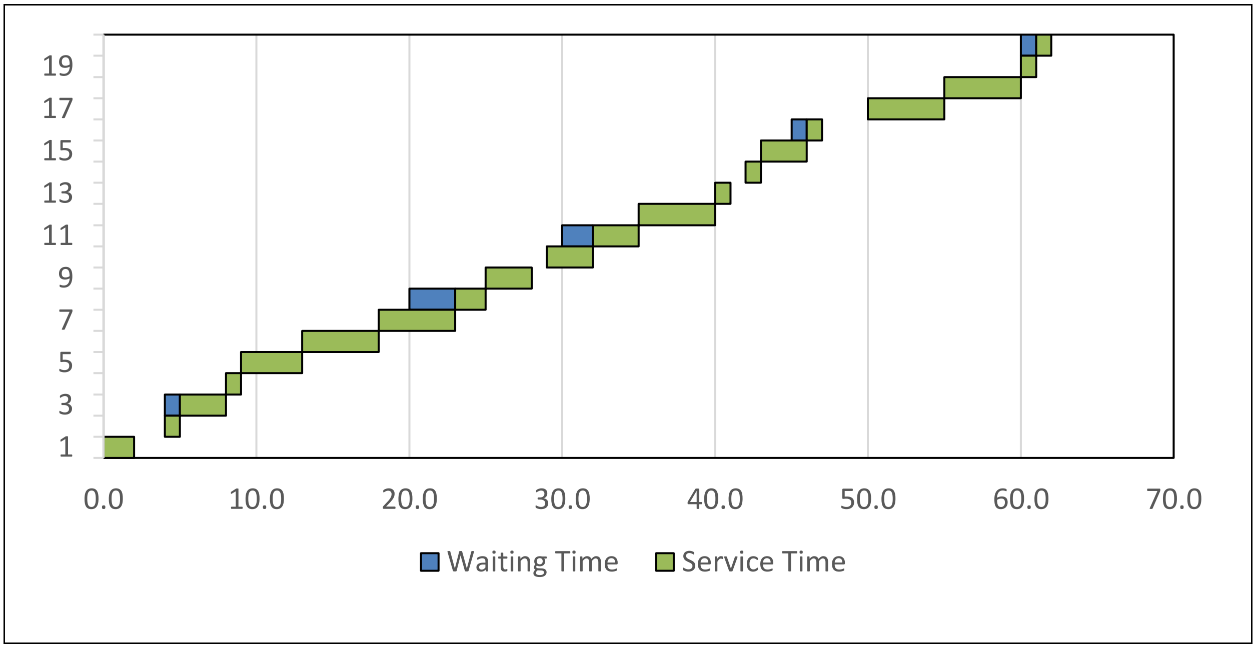


Figure 33. Sixth DES simulation run first 20 customers.

The seventh DES run, shown in Figure 34, indicated a significant difference between customers who would have occupied a VPQ (M = 6.80, SD = 2.81) and customers in the normal service

queue (M = 22.40, SD = 5.15); $t(28) = 10.30$, $p = .001$. An initial rise and then a reduction in wait times contributed to a loss of advanced reservations early in the simulation run. Another loss occurred towards the end of the simulation and the overall filled to total reservations ratio was .75 or a loss of 25%.

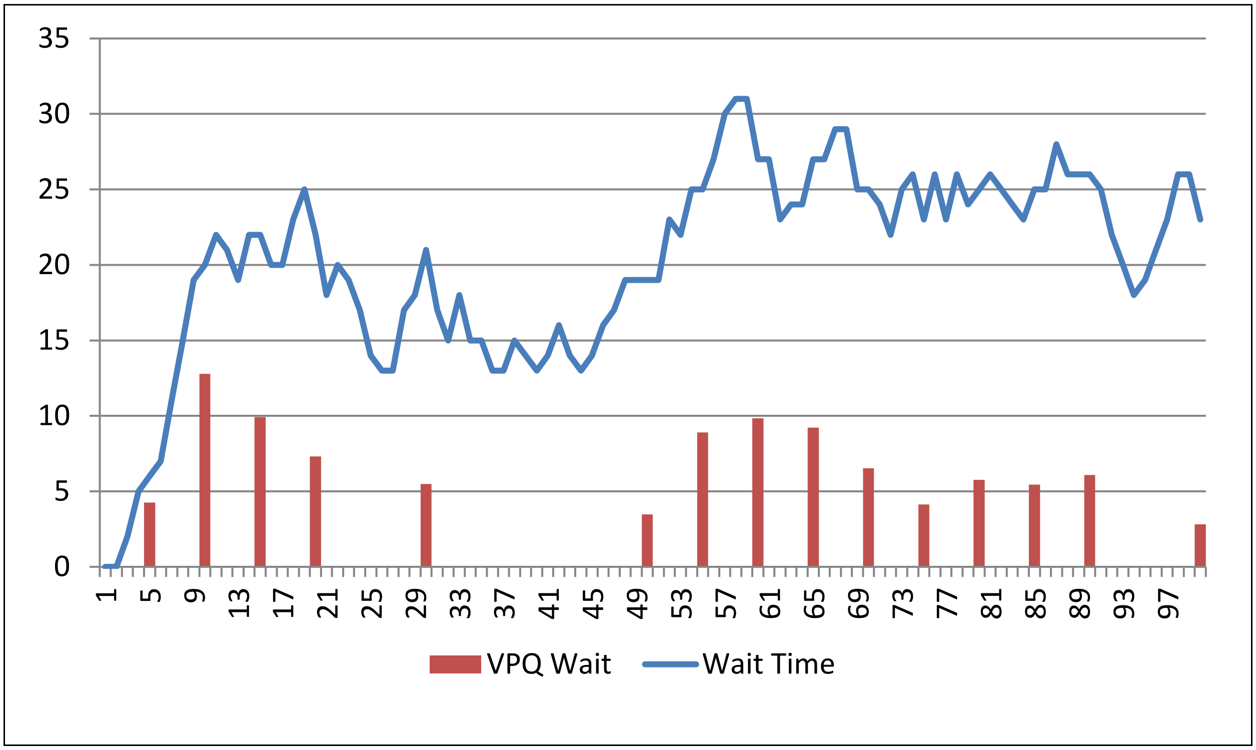


Figure 34. Seventh DES simulation run.

The bar graph for run number seven, shown in Figure 35, shows a rapid rise in wait times and the server idle for two time periods after the first arrival. The eighth DES run, shown in Figure 36, indicated a significant difference between customers who would have occupied a VPQ (M = 5.01, SD = 3.01) and customers in the normal service queue (M = 12.71, SD = 3.97); $t(26) = 5.79$, $p = .001$. Although this simulation run had the lowest filled to total reservations ratio of .70, or a 30% loss, it was still significant in reducing wait times between arrivals in the VPQ and normal queue.

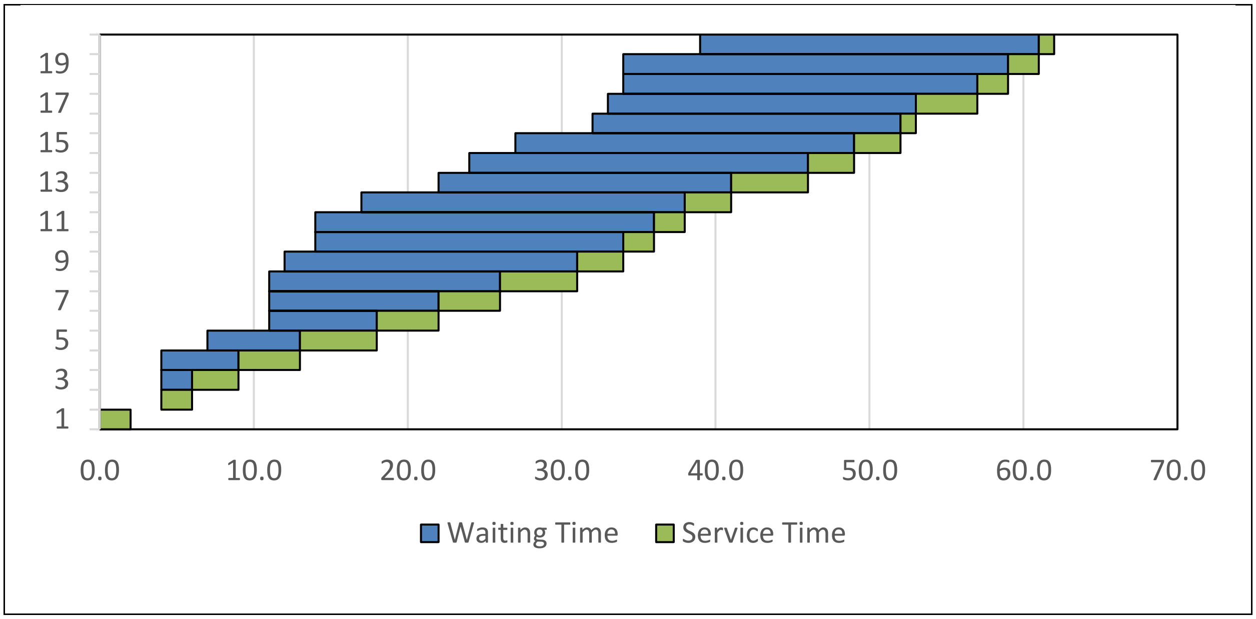


Figure 35. Seventh DES simulation run first 20 customers.

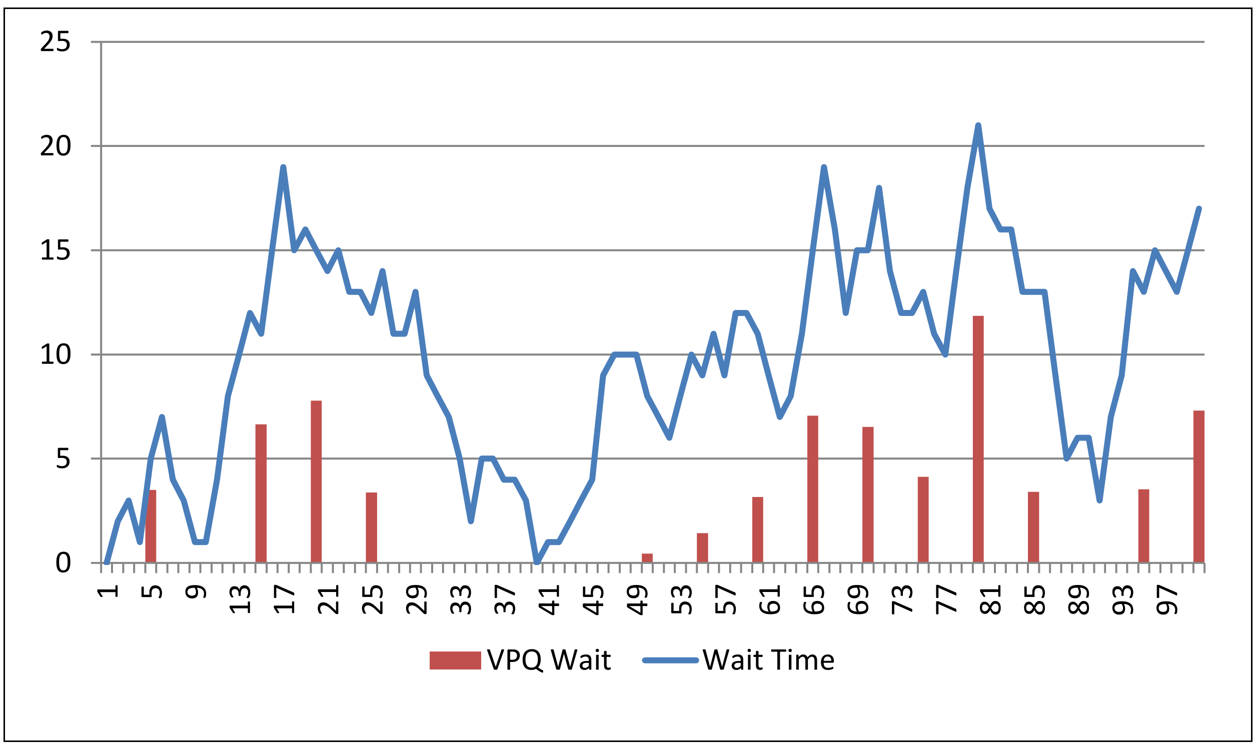


Figure 36. Eighth DES simulation run.

The ninth DES run, shown in Figure 37, indicated a significant difference between customers who would have occupied a VPQ (M = 10.95, SD = 5.27) and customers in the normal service

queue (M = 27.00, SD = 10.07); $t(36) = 6.15$, $p = .001$. This run had a filled to total reservations ration of .95 or 5% loss of advanced reservations.

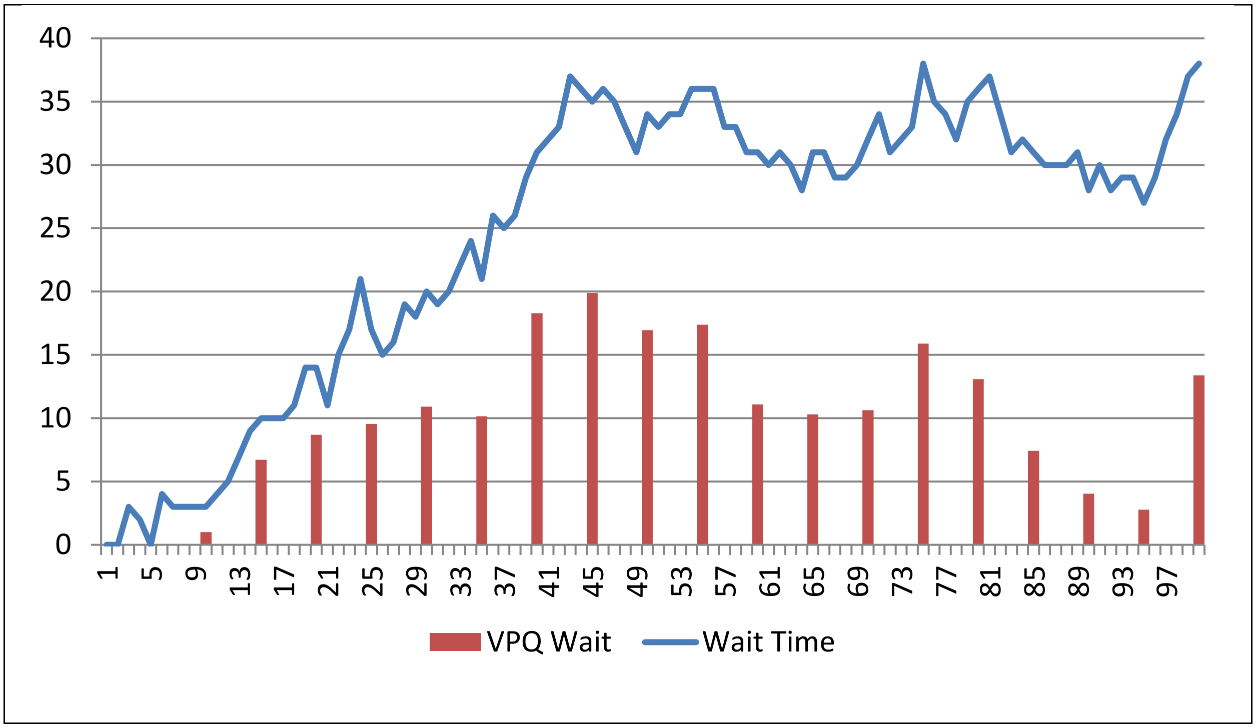


Figure 37. Ninth DES simulation run.

The final DES run, shown in Figure 38, indicated a significant difference between customers who would have occupied a VPQ (M = 15.33, SD = 5.25) and customers in the normal service queue (M = 46.55, SD = 15.12); $t(38) = 8.72$, $p = .001$. There were no advanced reservation losses and the filled to total reservations ratio was 1.00.

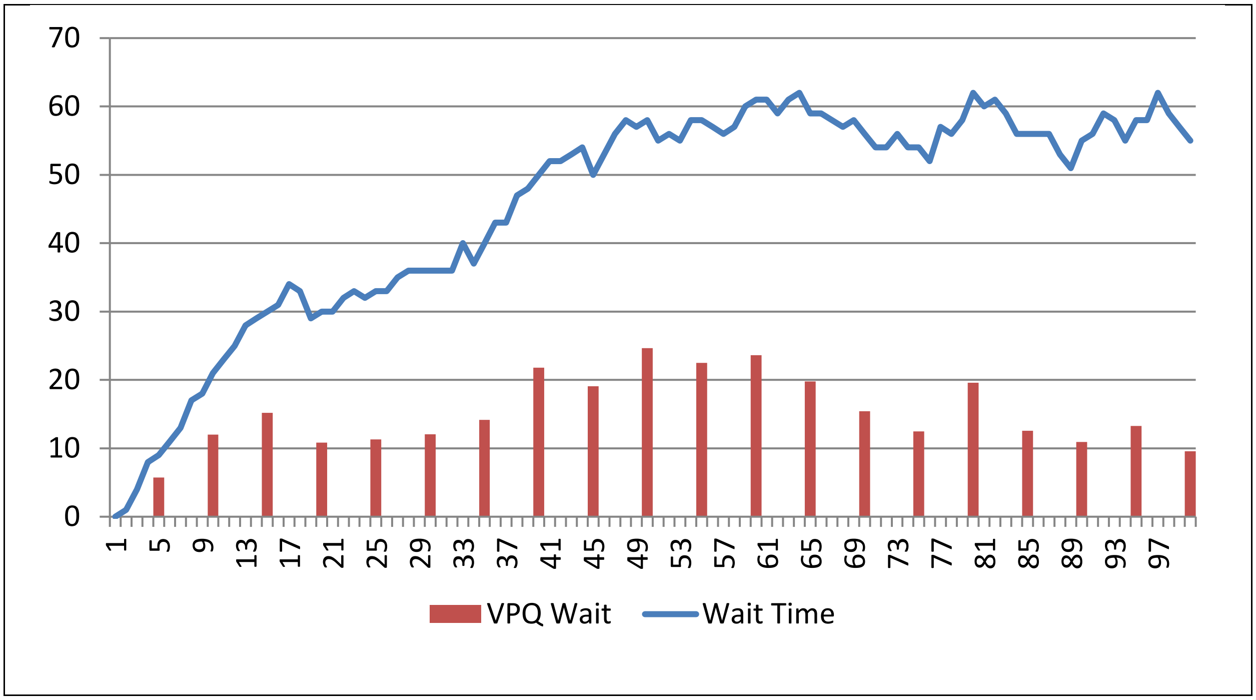


Figure 38. Tenth DES simulation run.

Summary

The results of this study conclude that perceived justice and patience threshold play a significant role in intent to abandon a service queue. A distinction was made between standing in a line and waiting in a call queue and respondents were asked a series of questions based on the respective scenarios. The exploratory instrument that was developed and used is promising in future research and further validation. It is noted that perceived justice (PJ) is also a factor in abandoning a call queue even though an individual may be totally unaware of others also in the queue. The intent of the including the survey as part of the investigation was to understand the impact of an individual's patience threshold on intent to abandon a service queue. Given that the proposed VPQ infrastructure may reduce the wait times for a select population in a call queue, a significant finding of that relationship would justify the potential benefits of a VPQ implementation.

The DES was successful in demonstrating significant time savings between a simulated individual in a VPQ advanced reservation and someone in the normal service queue. Ten simulation runs were conducted with varying degrees of significance, but overall successful in realizing the time savings that could be awarded to those chosen individuals. The supported hypotheses summary is shown in Table 5.

Table 5

Supported Hypothesis Summary

			t-value	P value
	Apparent (Line Standing) Queue			
H_1	Customers who abandon a call queue do so because their patience time has been exceeded.	Supported	6.268	.001
H_2	Customers who abandon a call queue do so because they feel a social injustice has occurred.	Supported	4.443	.001
	Phantom (Call) Queue			
H_1	Customers who abandon a call queue do so because their patience time has been exceeded.	Supported	6.470	.001
H_2	Customers who abandon a call queue do so because they feel a social injustice has occurred.	Supported	3.447	.001
H_3	$W_{VPQ} < W_{SQ}$			
	Run 1	Supported	4.607	.001
	Run 2	Supported	5.781	.001
	Run 3	Supported	2.742	.010
	Run 4	Supported	6.664	.001
	Run 5	Supported	5.588	.001
	Run 6	Supported	2.078	.046
	Run 7	Supported	10.297	.001
	Run 8	Supported	5.789	.001
	Run 9	Supported	6.153	.001
	Run 10	Supported	8.723	.001

Discussion and Limitations

Discussion

The primary purpose of this study was to evaluate the efficacy of the proposed VPQ through an experimental computer simulation of its operation. The intent was to evaluate actual wait time reduction for a select population entering the VPQ that would be chosen by the servicer. Additionally, an exploratory study was conducted to understand the roles patience thresholds and perceived justice play on an individual’s intent to abandon a service queue. Although a VPQ implementation is designed only for call queues, both line standing and call queues were investigated with an exploratory instrument that was developed to measure those relationships. The developed survey instrument is promising as a starting point for continued research into perceptions of justice and patience thresholds within call queues. The research attempted to answer the following questions:

- Do customers who abandon a queue do so because their patience time has been exceeded?
- Do customers who abandon a queue do so because they feel an injustice has occurred?
- During simulated operations, are call wait times statistically less in the VPQ than in the normal service queue?

Two structural models emerged from this investigation. One considers patience thresholds and perceived justice on intent to abandon call queues and the other the same for line standing queues. The perceived justice construct was developed from Rawls' (1971) theory of justice as that of justice as fairness. Rawls offers a thought experiment on what he describes as the *original position* where individuals choose a just form of society without knowing individual attributes of

skill, intelligence, or pedigree. Rawls' asserts this can only occur under the *veil of ignorance* where everyone would likely choose fair and equitable rules for all in a new society. A deliberate comparison was made between Rawls' concepts of original position and veil of ignorance with that of individuals entering a call service queue. Considering FIFO is a universally accepted and assumed policy for most service queues, individuals entering a call queue expect fair and equitable treatment when seeking service. Patience thresholds and perceived justice are found to be factors on an individual's intent to abandon a service queue regardless if it is an apparent (line standing) or phantom (call) queue. This was interesting to note since individuals in a call queue should have no apparent indication of how many occupy the queue or if a FIFO policy were violated.

The DES included in this study was a simple representation of a single-server queue with Poisson arrivals and exponential service times. The VPQ advanced reservation feature was invoked deterministically with every fifth entry to the queue and being filled based on the average of all previous normal service queue wait times. This simple simulation produced positive results for VPQ utility and reducing the actual wait time for some individuals.

The first two hypotheses were supported through the developed survey instrument and applied to both line standing and call queues:

H_1: Patience Threshold (PT) is negatively related to Intent to Abandon (IA) a service queue.

H_2: Perceived Justice (PJ) is negatively related to Intent to Abandon (IA) a service queue.

The third hypothesis was supported through 10 simulation runs but with a slight variation in level of significance:

H_3: $W_{VPQ} < W_{SQ}$

H_{3a}: $W_{VPQ} \geq W_{SQ}$

Limitations

This study was limited in not having a theoretical framework or instrument to use in investigating the research questions. The developed instrument underwent a pilot launch for refinement but theorized variables thought to have an impact on intent to abandon a service queue did not show significant relationships. Another limitation was relying on a single-source for respondents (MTurk), although previous research indicates that good quality data has been gathered (Buhrmester et al., 2011). For the DES portion, no attempt was made to apply predictive analytics to real-world datasets to optimize invoking a VPQ advanced reservation to reduce wait times to the lowest possible values for those individuals selected. The DES portion also did not explore asymptotic analysis of a simulated VPQ in operation.

Recommendations

Further technical development of the proposed VPQ would be warranted to reduce the call wait times for a select population determined by the servicer. Although a VPQ would not alleviate extended wait times for all who desire service, it would be beneficial for reducing the wait times for some and possibly lead to an increase of perceived quality of service (Taylor, 1994). A movement from managing the perceptions of an individual’s wait time to managing the actual wait time for some may lead to a better experience and overall satisfaction.

Conclusions

The development survey instrument designed to investigate patience thresholds and perceived justice on intent to abandon a service queue provided promising results in understanding those relationships in both line standing and call queues. The DES portion of this research provided a simple representation of VPQ functionality with significant reductions in wait times found.

Recommendations for Future Research

Future research into the constructs around intent to abandon a service queue should include patience thresholds and perceived justice. In addition, investigation into perceived value of waiting should be researched more. The data collected during this survey did not indicate perceived value as an antecedent to patience thresholds or actual intent to abandon. Another area that warrants additional study is the application of Rawls' (1971) theory of justice to phantom or call queues. Considering perceived justice was a factor in intent to abandon a call queue for service, understanding how an individual’s perception of an injustice occurring within a phantom queue should be interesting to research.

The developed research instrument should be applied to similar studies on both line standing and call queues. A core framework exists within the instrument that indicates an impact on intent to abandon a service queue and future refinement may be possible if the delivery were modified to engage individuals after being subjected to a waiting experience (Taylor, 1994).

References

Aguir, S., Karaesmen, F., Akşin, O. Z., & Chauvet, F. (2004). The impact of retrials on call center performance. *OR Spectrum*, *26*(3), 353–376. https://doi.org/10.1007/s00291-004-0165-7

Akhtar, S., & Latif, M. (2010). Exploiting simulation for call centre optimization. In *WCE 2010 - World Congress on Engineering 2010* (Vol. 3, pp. 2112–2117). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-79959858939&partnerID=40&md5=97a4a84107983f471adf088936994f71

Aksin, Z., & Masini, A. (2008). Effective strategies for internal outsourcing and offshoring of business services: An empirical investigation. *Journal of Operations Management*, *26*(2), 239–256. https://doi.org/10.1016/j.jom.2007.02.003

Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Managment Society*, *16*(6), 665–688. https://doi.org/10.3401/poms.

Alexander, M., MacLaren, A., O'Gorman, K., & White, C. (2012). Priority queues: Where social justice and equity collide. *Tourism Management*, *33*(4), 875–884. https://doi.org/10.1016/j.tourman.2011.09.009

Ali, L. F. (2010). *A call center simulation study : Comparing the reliability of cross-trained agents to specialized agents* (Master's thesis). Retrieved from http://trace.tennessee.edu/cgi/viewcontent.cgi?article=1653&context=utk_gradthes

Allon, G., & Hanany, E. (2012). Cutting in line: Social norms in queues. *Management Science, 58*(3), 493–506. https://doi.org/10.1287/xxxx.0000.0000

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J., & Ohlmann, J. W. (2013). *Quantitative methods for business*. Mason, OH: South-Western, Cengage Learning.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Andrews, D. (2013). *Why does the other line always move faster?* New York, NY:Workman Publishing Co., Inc.

Angus, I. (2001). An Introduction to Erlang B and Erlang C. *Telemanagement*, *187*, 6–8.

Atlason, J., Epelman, M. A., & Henderson, S. G. (2004). Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, *127*(1), 333–358. https://doi.org/10.1023/B

Babbie, E. (1990). *Survey Research Methods* (2nd ed.). Belmont, CA: Wadsworth Publishing Company

Bapat, V., & Pruitte, E. B. (1998). Using simulation in call centers. In *Simulation Conference Proceedings, 1998. Winter* (p. Vol. 2 pp. 1395–1399). https://doi.org/10.1007/s13398-014-0173-7.2

Bertoli, M., Casale, G., & Serazzi, G. (2009). JMT: performance engineering tools for system modeling. *ACM SIGMETRICS Performance Evaluation Review*, *36*(4), 10–15. https://doi.org/10.1145/1530873.1530877

Borst, S., Mandelbaum, A., & Reiman, M. I. (2004). Dimensioning large call centers. *Operations Research*, *52*(1), 17–34. https://doi.org/10.1287/opre.1030.0081

Bouzada, M. A. C. (2009). Dimensioning a call center: Simulation or queue theory? *Journal of Operations and Supply Chain Management*, *2*(2), 34–46.

Brady, F. N. (2002). Lining up for Star-Wars tickets: Some ruminations on ethics and economics based on an internet study of behavior in queues. *Journal of Business Ethics*, *38*(1–2), 157–165. https://doi.org/10.1023/A:1015725130366

Brockmeyer, E., Halstrom, H. L., & Jensen, A. (1948). *The life and works of A.K. Erlang*. Copenhagen, Denmark: Acta Polytechnical Scandinavica.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, *100*(469), 36–50. https://doi.org/10.1198/016214504000001808

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Cooper, D. R., & Schindler, P. S. (2011). *Business research methods* (7th ed.). New York, NY: McGraw-Hill Irwin.

Creswell, J. W. (2012). *Educational research planning, conducting, and evaluating quantitative and qualitative research* (3rd ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage.

De Mesquita, M. A., & Hernandez, A. E. (2006). Discrete event simulation of queues with spreadsheets: A teaching case. In *Proceedings of the 2006 Winter Simulation Conference* (pp. 2277–2283).

Deloitte. (2015). *2015 Global contact center survey: Executive summary*. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/us/Documents/process-and-operations/us-sdt-2015-executive-summary-for-gcc-survey.pdf

Demiriz, A., Kula, U., & Akbilek, N. (2009). A framework for balanced service and cross-selling by using queuing science. *Journal of Intelligent Manufacturing*, *20*(2), 249–257. https://doi.org/10.1007/s10845-008-0232-z

Eichhorn, B. R. (2014). Common method variance techniques. In *Cleveland State University, Department of Operations & Supply Chain Management*. Cleveland, OH. Retrieved from http://mwsug.org/2014-proceedings.html

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, *20*(6), 87–98.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Fukunaga, A., Hamilton, E., Fama, J., Andre, D., Matan, O., & Nourbakhsh, I. (2002). Staff scheduling for inbound call centers and customer contact centers. *Ai Magazine*, *23*(4), 30–40. Retrieved from http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=7&SID=2ArH9NWE5W3ujJxXWuO&page=1&doc=1

Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, *5*(2), 79–141. https://doi.org/10.1287/msom.5.2.79.16071

Gans, N., Liu, N., Mandelbaum, A., Shen, H., & Han Ye. (2010). Service times in call centers : Agent heterogeneity and learning with some operational consequences. *IMS Collections Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, *22*, 99–123. https://doi.org/10.1214/10-IMSCOLL608

Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, *4*(3), 208–227. https://doi.org/10.1287/msom.4.3.208.7753

Green, L. V., Kolesar, P. J., & Soares, J. (2003). An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management*, *12*(1), 46–61. https://doi.org/10.1111/j.1937-5956.2003.tb00197.x

Guo, P., & Zipkin, P. (2007). Analysis and comparison of queues with different levels of delay information. *Management Science*, *53*(6), 962–970. https://doi.org/10.1287/mnsc.1060.0686

Gurumurthi, S., & Benjaafar, S. (2004). Modeling and analysis of flexible queueing systems.

Naval Research Logistics, *51*(5), 755–782. https://doi.org/10.1002/nav.20020

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New Challenges to International Marketing* (Vol. 20, pp. 277–319). Emerald Group Publishing Limited. https://doi.org/10.1016/0167-8116(92)90003-4

Holman, D., Batt, R., & Holtgrewe, U. (2007). The global call center report : International perspectives on management and employment.

Huff, C., & Tingley, D. (2015). “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3). https://doi.org/10.1177/2053168015604648

Ibrahim, R., & L’Ecuyer, P. (2012). Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management*, *15*(1), 72–85. https://doi.org/10.1287/xxxx.0000.0000

Ipeirotis, P. G. (2010). Demographics of mechanical turk. https://doi.org/10.2139/ssrn.1585030

Jouini, O., Aksin, Z., & Dallery, Y. (2011). Call centers with delay Information: models and insights. *Manufacturing & Service Operations Management*, *13*(4), 534–548. https://doi.org/10.1287/msom.1110.0339

Jouini, O., Dallery, Y., & Aksin, Z. (2009). Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics*, *120*(2), 389–399. https://doi.org/10.1016/j.ijpe.2008.01.011

Koole, G., & Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, *113*(1–4), 41–59. https://doi.org/10.1023/A:1020949626017

Koole, G., & Pot, A. (2006). An overview of routing and staffing algorithms in multi-skill customer contact centers. *Submitted for Publication*, 1–42. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.1714&rep=rep1&type=pdf

Larson, R. C. (1987). OR forum - perspectives on queues: Social justice and the psychology of queuing. *Operations Research*, *35*(6), 895–905.

Larson, R. C. (1988). There's more to a line than its wait. *Technology Review*, *91*(5), 60–67.

Leong, T. Y. (2007). Simpler spreadsheet simulation of multi-server queues. *INFORMS Transactions on Education*, *7*(2), 172–177. https://doi.org/10.1287/ited.7.2.172

Lewis, B. G., Herbert, R. D., Summons, P. F., & Chivers, W. J. (2007). Agent-based simulation of a multi-queue emergency services call centre to evaluate resource allocation. In *MODSIM07 - Land, Water and Environmental Management: Integrated Systems for Sustainability, Proceedings* (pp. 11–17). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-67649188076&partnerID=40&md5=40173645281c8e7dc47c20f76f6c2d8a

Li, C., & Yue, D. (2015). A queuing model of the N-design multi-skill call center with impatient customers. *International Journal of Service, Science and Technology*, *8*(4), 51–60.

Maister, D. H. (2005). *The psychology of waiting lines*. Retrieved from http://www.davidmaister.com

Mandelbaum, A., & Momčilović, P. (2015). Personalized queues : The customer view , via a fluid model of serving least-patient first. *Queueing Systems*, *87*(1–2), 23–53.

Mandelbaum, A., & Shimkin, N. (2000). A model for rational abandonments from invisible queues. *Queuing Systems*, *36*(1), 141–173. https://doi.org/10.1023/A:1019131203242

Mandelbaum, A., & Zeltyn, S. (2007). Service engineering in action: The palm/Erlang-A queue, with applications to call centers. *Advances in Services Innovations*, 17–45. https://doi.org/10.1007/978-3-540-29860-1_2

Mandelbaum, A., & Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, *57*(5), 1189–1205. https://doi.org/10.1287/opre.1080.0651

Massey, W. A. (2002). The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, *21*(2), 173–204. https://doi.org/10.1023/A:1020990313587

Mathew, B., & Nambiar, M. K. (2013). A tutorial on modelling call centres using discrete event simulation. *Proceedings - 27th European Conference on Modelling and Simulation, ECMS 2013*, 315–321. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84900328109&partnerID=tZOtx3y1

Mehrotra, V., & Fama, J. (2003). Call center simulations: methods, challenges, and opportunities. In *Proceedings of the 35th conference on winter simulation* (pp. 135–143). Retrieved from http://portal.acm.org/citation.cfm?id=1030818.1030837

Newell, G. F. (1982). *Applications of queuing theory* (2nd ed.). Cambridge, England: University

Press.

Paek, T., & Horvitz, E. (2004). Optimizing Automated Call Routing by Integrating Spoken Dialog Models with Queuing Models. In *HLT-NAACL* (pp. 41–48).

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419. https://doi.org/10.2139/ssrn.1626226

Perry, R., & Zarsky, T. Z. (2013). Queues in law. In *Iowa Law Review* (Vol. 99, p. 1595).

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research : A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Ringle, C. M., Wende, S., & Will, A. (2005). SmartPLS. Hamburg, Germany. Retrieved from http://www.smartpls.de/

Robbins, T. R., Medeiros, D. J., & Harrison, T. P. (2010). Does the Erlang C model fit in real call centers? In *Proceedings of the 2010 Winter Simulation Conference* (pp. 2853–2864). https://doi.org/10.1109/WSC.2010.5678980

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers: Shifting demographics in mechanical turk. In *CHI 2010* (pp. 2863–2872). https://doi.org/10.1145/1753846.1753873

Sankaranarayanan, K. (2011). *Study on behavioral patterns in queuing : Agent based modeling and experimental approach* (Doctoral dissertation, University of Lugano). Retrieved from https://doc.rero.ch/record/27284/files/2011ECO006.pdf

Sekaran, U., & Bougie, R. (2013). *Research methods for business: A skill building approach* (6th ed.). West Sussex, England: John Wiley & Sons.

Shen, H. (2010). Statistical analysis of call-center operational data: Forecasting call arrivals, and analyzing customer patience and agent service. In *Wiley Encyclopedia of Operations Research and Management Science* (pp. 1–10). https://doi.org/10.1002/9780470400531.eorms0827

Sorokin, V. (1988). *The Queue*. New York, NY: Readers International.

Taylor, S. (1994). Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing*, *58*(2), 56–69. https://doi.org/10.2307/1252269

Templeton, G. F., & Burney, L. L. (2017). Using a Two-Step Transformation to Address Non-Normality from a Business Value of Information Technology Perspective. *Journal of*

Information Systems, *31*(2), 149–164. https://doi.org/10.2308/isys-51510

Wallace, R. B., & Whitt, W. (2005). A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, *7*(4), 276–294. https://doi.org/10.1287/msom.1050.0086

Weinberg, J., Brown, L. D., & Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, *102*(480), 1185–1198. https://doi.org/10.1198/016214506000001455

Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science*, *51*(2), 221–235. https://doi.org/10.1287/mnsc.1040.0302

Winston, W. L. (1994). *Operations research: Applications and algorithms* (3rd ed.). Belmont, CA: Duxbury Press.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis : Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1993). The nature and determinants of customer expectations of service. *Journal of the Academy of Marketing Science*, *21*(1), 1–12. https://doi.org/10.1177/0092070393211001

Zeltyn, S., & Mandelbaum, A. (2005). Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems*, *51*(3–4 SPEC. ISS.), 361–402. https://doi.org/10.1007/s11134-005-3699-8

Zhou, R., & Soman, D. (2003). Looking back: Exploring the psychology of queuing and the effect of the number of people behind. *Journal of Consumer Research*, *29*(4), 517–530. https://doi.org/10.1086/346247

Zohar, E., Mandelbaum, A., & Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, *48*(4), 566–583.

Zukerman, M. (2008). *Introduction to queueing theory and stochastic teletraffic models*. *Networks*. Retrieved from http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf

Appendices

Appendix A

UNIVERSITY OF THE INCARNATE WORD

July 3 2017

PI: Mr. William Pugh

Protocol title: Queue perceptions of social justice

William:

Your request to conduct the study titled "Queue perceptions of social justice" was approved by Exempt review on 07/03/2017. Your IRB approval number is 17-07-001. Any written communication with potential subjects or subjects must be approved and include the IRB approval number.

Please keep in mind these additional IRB requirements:

- This approval will expire **one year** from 07/03/2017.
- Request for continuing review must be completed for projects extending past one year. Use the **IRB Continuing Review Request form**.
- Changes in protocol procedures must be approved by the IRB prior to implementation except when necessary to eliminate apparent immediate hazards to the subjects. Use the **IRB Amendment Request form**.
- Any unanticipated problems involving risks to subjects or others must be reported immediately.

Approved protocols are filed by their number. Please refer to this number when communicating about this protocol.

Approval may be suspended or terminated if there is evidence of a) noncompliance with federal regulations or university policy or b) any aberration from the current, approved protocol.

Congratulations and best wishes for successful completion of your research. If you need any assistance, please contact the UIW IRB representative for your college/school or the Office of Research Development.

Sincerely,

Ana Wandless-Hagendorf, PhD, CPRA

Ana Wandless-Hagendorf, PhD, CPRA

Research Officer, Office of Research Development

University of the Incarnate Word

(210) 805-3036

wandless@uiwtx.edu

Appendix B

Consumer Perceptions of Service Queues Survey

Waiting in a line, or queue, for service is a common activity for many. Service queues can take on many forms. Among the more common types are physical queues, where one waits in a line with other customers and call queues, where one waits on the phone to receive a particular service. You will be presented two situations where waiting in a queue is required. You will be asked to answer a serious of questions based on each queue scenario.

INFORMED CONSENT: You are being asked to take part in a research study regarding consumer perceptions of waiting in queues. This survey should take 15 minutes or less to complete. Participation in the survey is voluntary and will not cause you discomfort or inconvenience. No identifying information is being requested and your IP address is not being retained as part of your response to this survey. Participation in this study is completely voluntary. Please be aware that if you decide to participate, you may stop participating at any time. If you have questions or you wish to report a problem that may be related to this study, contact William Pugh, wpugh@uiwtx.edu. For questions about your rights as a research participant or to discuss problems, complaints or concerns about a research study, or to obtain information or offer input, contact the University of the Incarnate Word (UIW) Institutional Review Board (IRB) at (210) 805-3036. This research and survey tool has been approved by the UIW IRB (IRB #17-07-001).

By agreeing to these conditions, you are indicating that you have read the description of the study, are over the age of 18, and that you agree to the terms as described.

1. I will wait in a queue if I believe the service is extremely valuable to me.

2. I feel service should be provided within the announced wait times.

3. I have very little patience for waiting in a queue for service.

4. The longer I wait in a queue for service, the more likely I am to abandon the queue.

5. When I must wait in a queue to receive service, I feel there should be an opportunity for me to express a priority need so I can receive service faster.

6. I will wait in a queue for service even if I am delayed for other activities.

7. I feel service should be provided to those who have waited the longest.

8. I avoid situations where I must wait in a queue for service for an extended period of time.

9. I will abandon a service queue if I feel others are served before me who have not waited as long as I have.

10. When I must wait in a queue to receive service, I feel my needs are more important than the needs of others.

11. I will wait in a queue even if I feel the service is of little value.

12. When I must wait in a queue for service, I feel others waiting in the queue expect to receive service in the order of arrival.

13. I feel the expected wait time should always be provided while I am in a queue for service.

14. I abandon a queue if I perceive others are being served before me.

15. When I must wait in a queue to receive service, I feel it would be best to randomly select individuals in the queue.

16. I will wait in a queue even if I can receive the same service through another channel such as a website, sending a text or email.

17. When I must wait in a queue for service, I expect everyone to be served in the order of their arrival.

18. I feel waiting in a queue is usually required in order to receive service.

19. As a valued customer, I feel there should be no waiting for service.

20. I prefer to make an advanced reservation for situations where I must wait in a queue for service.

21. When I must wait in a queue to receive service, I feel it is fair that service is provided first to those who have waited longer.

22. I will consider abandoning a queue for service if I must wait for any amount of time.

23. I will abandon a service queue if the wait time exceeds my expectation.

24. I will wait in a queue if I believe the service is extremely valuable to me.

25. I feel service should be provided within the announced wait times.

26. I have very little patience for waiting in a queue for service.

27. The longer I wait in a queue for service, the more likely I am to abandon the queue.

28. When I must wait in a queue to receive service, I feel there should be an opportunity for me to express a priority need so I can receive service faster.

29. I will wait in a queue for service even if I am delayed for other activities.

30. I feel service should be provided to those who have waited the longest.

31. I avoid situations where I must wait in a queue for service for an extended period of time.

32. I will abandon a service queue if I feel others are served before me who have not waited as long as I have.

33. When I must wait in a queue to receive service, I feel my needs are more important than the needs of others.

34. I will wait in a queue even if I feel the service is of little value.

35. When I must wait in a queue for service, I feel others waiting in the queue expect to receive service in the order of arrival.

36. I feel the expected wait time should always be provided while I am in a queue for service.

37. I abandon a queue if I perceive others are being served before me.

38. When I must wait in a queue to receive service, I feel it would be best to randomly select individuals in the queue for service.

39. I will wait in a queue even if I can receive the same service through another channel such as a website, sending a text or email.

40. When I must wait in a queue for service, I expect everyone to be served in the order of their arrival.

41. I feel waiting in a queue is usually required in order to receive service.

42. As a valued customer, I feel there should be no waiting for service.

43. I prefer to make an advanced reservation for situations where I must wait in a queue for service.

44. When I must wait in a queue to receive service, I feel it is fair that service is provided first to those who have waited longer.

45. I will consider abandoning a queue for service if I must wait for any amount of time.

46. I will leave a service queue if the wait time exceeds my expectation.