Queueing Models of Call Centers: An Introduction

Ger Koole

Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Avishai Mandelbaum [*]

Industrial Engineering and Management, Technion, Haifa 32000, Israel

This is a survey of some academic research on telephone call centers. The surveyed research has its origin in, or is related to, queueing theory. Indeed, the "queueing-view" of call centers is both natural and useful. Accordingly, queueing models have served as prevalent standard support tools for call center management. However, the modern call center is a complex socio-technical system. It thus enjoys central features that challenge existing queueing theory to its limits, and beyond.

The present document is an abridged version of a survey that can be downloaded from `www.cs.vu.nl/obp/callcenters` and `ie.technion.ac.il/∼serveng`.

Keywords: call centers, queueing models

1. Introduction

Call centers, or their contemporary successors contact centers, are the preferred and prevalent way for many companies to communicate with their customers. The call center industry is thus vast, and rapidly expanding in terms of both workforce and economic scope. For example, it is estimated that 3% of the U.S. and U.K. workforce is involved with call centers, the call center industry enjoys a annual growth rate of 20% and, overall, more than half of the business transactions are conducted over the phone. (See `callcenternews.com/resources/statistics.shtml` for a collection of call center statistics.)

Within our service-driven economy, telephone services are unparalleled in scope, service quality and operational efficiency. Indeed, in a large best-practice call center, many hundreds of agents could cater to many thousands of phone callers per hour; agents utilization levels could *average* between 90% to 95%; no customer encounters

[*] Research partially supported by the ISF (Israeli Science Foundation) grant 388/99-02, by the Technion funds for the promotion of research and sponsored research, and by Whartons' Financial Institutions Center

a busy signal and, in fact, about half of the customers are answered *immediately*; the waiting time of those delayed is measured in seconds, and the fraction that abandon while waiting varies from the negligible to mere 1-2% (e.g., see Figures 2 and 3). The design of such an operation, and the management of its performance, surely must be based on sound scientific principles. This is manifested by a growing body of academic multi-disciplinary research, devoted to call centers, and ranging from Mathematics and Statistics, through Operations Research, Industrial Engineering, Information Technology and Human Resource Management, all the way to Psychology and Sociology. (The bibliography [35] covers over 200 research papers.) *Our goal here is to survey part of this literature, specifically that which is based on mathematical queueing models and which potentially supports Operations Research and Management.*

1.1. What is a call center?

A *call center* constitutes a set of resources (typically personnel, computers and telecommunication equipment), which enable the delivery of services via the telephone. The working environment of a large call center could be envisioned as an endless room with numerous open-space cubicles, in which people with earphones sit in front of computer terminals, providing tele-services to unseen customers. Most call centers also support Interactive Voice Response (IVR) units, also called Voice Response Units (VRU's), which are the industrial versions of answering machines, including the possibilities of interactions. But more generally, a current trend is the extension of the call center into a *contact center*. The latter is a call center in which the traditional telephone service is enhanced by some additional multi-media customer-contact channels, commonly VRU, e.mail, fax, Internet or chat (in that order of prevalence).

Most major companies have reengineered their communication with customers via one or more call centers, either internally-managed or outsourced. The trend towards contact centers has been stimulated by the societal hype surrounding the Internet, by customer demand for channel variety, and by acknowledged potential for efficiency gains.

1.2. Technology

The large-scale emergence of call centers, noticeably during the last decade, has been enabled by technological advances in the area of Information and Communication Technology (ICT). First came PABX's (Private Automatic Branch Exchanges, or simply PBX), which are the telephone exchanges within companies. A PABX connects, via trunks (telephone lines), the public telephone network to telephones within the call centers. These, in turn, are staffed by telephone *agents*, often called CSR's for Cusomer Service Representatives, or simply "rep's" for short. Intermediary between the PABX and the agents is the ACD (Automatic Call Distribution) switch, whose role is to distribute calls among idle qualified agents. A secondary responsibility of the ACD is the archival collection of operational data, which is of prime importance as far as call center research is concerned. While there exists a vast telecommunications literature on the physics of telephone-traffic and the hardware (technology) of call centers, *our survey focuses on the service contact between customers and agents*, sometimes referred to as the service's "moment of truth".

Advances in information technology have contributed as importantly as telecommunication to the accelerated evolution of call centers. To wit, rather than search for a paper file in a central archive, that renders impossible an immediate or even fast handling of a task related to that file, nowadays an agent can access, almost instantaneously, the needed file in the company's data base. A new trends in ICT is the access of customer files in an automatic way. The relevant technology is CTI (Computer Telephony Integration), which does exactly what its name suggests. In fact, this can go further. Consider, for example, a customer who seeks technical support from a telephone help-desk. That customer can be often automatically identified by the PABX, using ANI (Automatic Number Identification). This triggers the CTI to search for the customer's history file; information from the file then *pops up* on the agent's computer screen, detailing all potentially relevant support for the present transaction, as well as pointers for likely responses to the support request. Having identified the customer's need, this could all culminate in an almost instantaneous automatic e.mail or fax that resolves the customer's problem. In a business setting, CTI and ANI are used to identify, for example, cross- or up-selling opportunities and, hence, routing of the call to an appropriately skilled agent.

1.3. The world of call centers

Call centers can be categorized along many dimensions: functionality (help desk, emergency, tele-marketing, information providers, etc.), size (from a few to several thousands of agent seats), geography (single- vs. multi-location), agents charateristics (low-skilled vs. highly-trained, single- vs. multi-skilled), and more. A central characteristic of a call center is whether it handles *inbound* vs. *outbound* traffic. (Synonyms for inbound/outbound are incoming/outgoing.) *Our focus here is on inbound call centers*, with some attention given to mixed operations that blend in- and out-going calls. An example of such blending is when agents are utilizing their idle time to call customers that left IVR requests to be contacted, or customers that abandoned (and had been identified by ANI) to check on their wishes. Pure outbound call centers are typically used for advertisement or surveys - they will be only briefly described (and contrasted with pure inbound and mixed operations) in Subsection 3.5.

Modern call/contact centers however are challenged with multitude types of calls, coming in over different communication channels (telephone, internet, fax, e.mail., chat, mobile devices, etc.); agents have the skill to handle one or more types of calls (e.g., they can provide technical support for several products in several languages by telephone, e.mail or chat). Furthermore, the organizational architecture of the modern call center varies from the very flat, where essentially all agents are exposed to external calls, to the multi-layered, where a layer represents say a level of expertise and customers could potentially be transferred through several layers until being served to satisfaction. Further yet, a call center could in fact be the virtual embodiment of few-to-many geographically dispersed call centers (from the very large, connected over several continents - for example, mid-West U.S.A. with Ireland and India - to the very small, constituting individual agents that work from their homes in their spare time).

1.4. Management and quality of service

There exists a large body of literature on the management of call centers, both in the academia (Section VII in [35] contains close to 50 references) and even more so in the trade literature.

Typically, call center goals are formulated as the provision of service at a given quality, subject to a specified budget (more on this momentarily). While Service Quality is a very complicated notion, to which numerous articles and books have been devoted [25,9,21], a highly simplified approach suffices for our purposes. We measure service quality along two dimensions: qualitative (psychological) and quantitative (operational). The former relates to the way in which service is provided and perceived (am I satisfied with the answer, is the agent friendly, etc.; for example, [49]). The latter relates more to service accessibility (how long did I have to wait for an answer, was I forced into calling back, etc.). Models in support of the qualitative aspects of service quality are typically empirical, originating in the Social Sciences or Marketing (see Sections III, IV and VIII in [35]). *Models in support of quantitative management are typically analytical, and here we focus on the subset of such models that originates in Operations Research in general and Queueing Theory in particular.*

Common practice is that upper management decides on the desired service level and then call center managers are called on to defend their budget. Similarly, costs can be associated with service levels (eg. toll-free services pay out-of-pocket for their customers' waiting), and the goal is to minimize total costs. These two approaches are articulated in [11]. It occurs, however, that profit can be linked directly to each individual call, for example in sales/mail-order companies. Then a direct trade-off can be made between service level and costs so as to maximizes overall profit. Two papers in which this is done are [4] and [2]. *In what follows we concentrate on the service level vs. cost (efficiency) trade-off.* The fact that salaries account for 60–70% of the total operating costs of a call center *justifies our looking mostly at personnel costs.* This is also the approach adopted by *workforce management tools*, that are used on a large scale in call centers. By concentrating on personnel, one presumes that other resources (such as ICT) are not bottlenecks (see however the work of [1,2]).

1.5. Performance measures

Operational service level is typically quantified in terms of some congestion or performance measures. Our experience, backed up by [21], suggests a *focus on abandonment, waiting and/or retrials*, which underscores the natural fit between queueing models and call centers (Subsection 1.7).

Performance measures are of course intercorrelated - see [50] for the remarkably linear relation between the fraction of abandoning customers and average waiting time. They could also convey more information that actually meets the eye. For example, in contrast to waiting statistics which are objective, abandonment and retrial measures are *subjective* in that they incorporates customers' view on whether the offered service is worth its wait (abandonment) or returning to (retrials). As another example, it turns out that one can quantify customers' patience in terms of the ratio between the fraction abandoned to the fraction served - indeed, it is shown in [39] that this ratio can be also interpreted as that between the average time that customers are *willing* to wait to the average time that they *expect* to wait.

For performance measures to be useful, they must be archived at a proper resolution and observed at the appropriate frequency. Ideally, one would like to store, for each individual transaction at the call center, its operational and business characteristics. This raw data can then be mined for exploratory purposes, or aggregated into performance measures for management use. For example, Figure 3 exhibits the prevailing standard, under which operational data is averaged over half-hour intervals. Such an averaging, however, is insufficient for deeper needs, as amply demonstrated in [39].

1.6. A scientific approach to management

In the practice of call center management, a quantitative approach often amounts to merely monitoring performance and intervening if that is considered necessary. The call center manager tracks performance indicators and *reacts* when they reach unacceptable levels; for example, too many customers are waiting or too many agents are idle. These reactions are typically based on subjectively-biased experiences, and a decision is doomed "poor" or "wrong" if the resulting performance turns out worse than expected.

In a more scientific approach, management is pro-active rather than reactive - for example, ensuring that waiting is scarce rather than adding agents when waiting becomes excessive. Here quantitative models - analytical or simulation - turn out useful for developing rules-of-thumb and intuition, or practically supporting design and control.

For example, the “what-if” scenarios in the Introduction to [11] demonstrate, via a simple analytical model, that call centers are typically extremely sensitive to changes in underlying parameters; this is closely related to the square-root principle for staffing, which is a rule-of-thumb that is presented below. Models have in fact become integral parts of the widely used workforce scheduling tools; but such uses rarely go beyond the rudimentary M/M/s (Erlang-C) queue, let alone the more sophisticated models that are surveyed in Section 3.

1.7. Queueing Theory and Science

Queues in service operations are often the arena where customers, service providers (servers, or agents) and managers establish contact, in order to jointly create the service experience. Process-wise, queues play in services much the same role as inventories in manufacturing. But in addition, “human queues” express preferences, complain, abandon and even spread around negative impressions. Thus, *customers* treat the queueing experience as a window to the service-providing party, through which their judgement of it is shaped for better or worse. *Managers* can use queues as indicators (queues are the means, not the goals) for control and improvement opportunities. Indeed, queues provide unbiased quantifiable measures (these are not abundant in services), in terms of which performance is relatively easy to monitor and goals are naturally formulated.

Research in quantitative call center management is concerned with the development of scientifically-based design principles and tools (often culminating in software), that support and balance service quality and efficiency, from the likely conflicting perspectives of customers, servers, managers, and often also society. Queueing models constitute a natural convenient nurturing ground for the development of such principles and tools [24,11]. However, the existing supporting (Queueing) theory has been somewhat lacking, as will now be explained.

The bulk of what is called Queueing Theory, consists of research papers that formulate and analyze queueing models with a realistic flavor. Most papers are knowledge-driven, where “solutions in search of a problem” are developed. Other papers are problem-driven, but most do not go far enough in the direction of a practical solution. Only some articles develop theory that is either rooted in or actually settles a real-world problem, and scarcely few carry the work as far as validating the model or the solution [26,29]. In concert with this state of affairs, not much is available of what could be called *Queueing Science*, or perhaps the Science of Congestion, which should supplement traditional queueing theory with empirically-based models [50], observations [39] and experiments [45,34]. In call centers, and more generally service networks, such “Science” is lagging behind that in telecommunications, computers, transportation and manufacturing. Key reasons for the gap seem to be the difficulty of measuring service op-

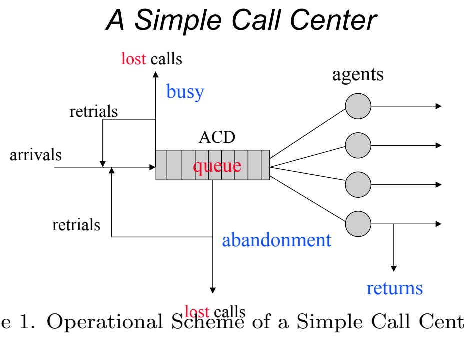


Figure 1. Operational Scheme of a Simple Call Center.

erations (see Section 2), combined with the need to incorporate human factors (which are notoriously difficult to quantify) - see Subsection 3.2 for a discussion of human patience while waiting in *tele-queues*.

1.8. Call centers as queueing systems

Call centers can be viewed, naturally and usefully, as queueing systems. This comes clearly out of Figure 1, which is an operational scheme of a simple call center. (See Subsection 3.1 for an elaboration.)

In a queueing model of a call center, the customers are callers, servers (resources) are telephone agents (operators) or communication equipment, and tele-queues consist of callers that await service by a system resource. The simplest and most-widely used such model is the $M/M/s$ queue, also known in call center circles as Erlang C. For most applications, however, Erlang C is an over-simplification: for example, it assumes out busy signals, customers impatience and services spanned over multiple visits. These features are captured in Figure 1, which depicts a single finite-queue with abandonment [24] and retrials [48,29]. But the modern call center is often a much more complicated queueing *network*: even the mere incorporation of an IVR, prior to joining the agents' tele-queue, already creates two stations in tandem [15], not to mention having multiple teams of specialized or cross-trained agents [23,10], that are geographically dispersed over multiple interconnected call centers [32], and who are faced with time-varying loads [38] of calls by multi-type customers [5,2].

1.9. Keeping up-to-date

A fairly complete list of academic publications on call centers has been compiled in [35]. There are over 200 publications, arranged chronologically within subjects, each with its title and authors, source, full abstract and keywords. Given the speed at which call center technology and research are evolving, advances are perhaps best followed through the Internet, for example using a search engine.

2. Data

Any modeling study of call centers must necessarily start with a careful data analysis. For example, the simplest Erlang C queueing model of a call center requires the estimation of calling rate and mean service (holding) times. Moreover, the performance of call centers in peak hours is extremely sensitive to changes in its underlying parameters. (See Figure 3, and the discussion in Subsection 3.2.) It follows that an extremely accurate estimation/forecasting of parameters is a prerequisite for a consistent service level and an efficient operation.

Section II in [35] lists only 16 papers on the statistics and forecasting of call center data. Given the data-intensive hi-tech environment of modern call centers, combined with the importance of accurate estimation, it is surprising, perhaps astonishing, that so little research is available and so much is yet needed. (Compare this state-of-affairs with that of Internet and telecommunication - here, only few year ago, a fundamental change in the research agenda was forced on by data analysis, which revealed new phenomenon, for example heavy-tails and long-range-dependence.)

There is a vast literature on statistical inference and forecasting, but surprisingly little has been devoted to stochastic processes and much less to queueing models in general and call centers in particular (see Section II in [35] for some exceptions). Indeed, the practice of statistics and time series in the world of call centers is still at its infancy, and serious research is required to bring it to par with its needs.

We distinguish between three types of call center data: operational, marketing, and psychological. *Operational* data is typically collected by the Automatic Call Distributor (ACD), which is part of the telephony-switch infrastructure (typically hardware-, but recently more and more software-based). *Marketing* or *Business* data is gathered by the Computer Telephony Integration/Information (CTI) software, that connects the telephony-switch with company data-bases, typically customer profiles and business histories. Finally, *psychological* data is deduced from surveys of customers, agents or managers. It records subjective perceptions of service level and working environment, and will not be discussed here further.

Existing performance models are based on operational ACD data. The ultimate goal, however, is to integrate data from the three sources mentioned above, which is essential if one is to understand and quantify the role of (operational) service-quality as a driver for business success.

3. Performance models

The essence of *operations management* in a call center is the matching of service requests (demand) with resources (supply). The fundamental tradeoff is between service quality vs. operational efficiency. *Performance analysis* supports this tradeoff by calculating attained service level and resource occupancy/utilization as functions of traffic load and available resources. We start with describing the simplest such models and then expand to capture main characteristics of today's highly complex contact centers.

3.1. Single-type customers and single-skill agents

A schematic operational model of a *simple* call center is depicted in Figure 1. The connotation is that of the old-times switch board, either those operated by telephone companies or as part of individual organizations, where telephone operators were connecting incoming calls physically to the proper extension/line. (Old papers on telephone services, as the classical Erlang [18] and Palm [41], were in fact modeling such switch boards.) Modern technology has now replaced these human operators by the ACD, that routes customers calls to idle agents. What renders the operation depicted above, as well as its model, "simple" is that there is a single type of calls that can be handled by all agents (statistically identical customers and servers).

The simplest and most used performance model is the stationary $M/M/s$ queue. It describes a single-type single-skill call center with s agents, operating over a short enough time-period so that calls arrive at a constant rate, yet randomly (Poisson); staffing level and service rates are also taken constant. The assumed stationarity could be problematic if the system does not relax fast enough, for example due to events such as an advertisement campaign or a mew-product release. The model assumes out busy signals, abandonment, retrials and time-varying conditions.

The reason for using the $M/M/s$ queue is of course the fact that there exist closed form expressions for most of its performance measures. However, $M/M/s$ predictions could turn out highly inaccurate because reality often "violates" its underlying assumptions, and these violations are not straightforward to model. For example, non-exponential service times leads one to the $M/G/s$ queue which, in stark contrast to $M/M/s$, is analytically intractable. One must then resort to approximations, out of which it turns out that service time affects performance through its coefficient-of-variation $C = E/\sigma$). Performance deteriorates (improves) as stochastic variability in service times increases (decreases). An empirical comparison between $M/M/s$ and $M/G/s$ models can be found in [48].

When modeling call centers, the useful approximations are typically those in heavy-traffic, namely high agents' utilization levels at peak hours. Consider again the $M/G/s$ queue. For small to moderate number of agents s, Kingman's classical result asserts that Waiting Time is approximately exponential, with mean as given above. Large s, on the other hand, gives rise to a different asymptotic behavior. This was first discovered by Halfin and Whitt [28] for the $M/M/s$ queue, and recently extended to $M/PH/s$ in [43]. We now discuss these issues within the context of two key challenges for call center management: agent staffing and economies of scale.

Square-root safety staffing The *square-root safety-staffing principle*, introduced formally in [11] but having existed long before, recommends a number of servers s given by

$$s \;=\; R+\Delta \;=\; R+\beta\sqrt{R}\;,\quad -\infty<\beta<\infty\;,$$

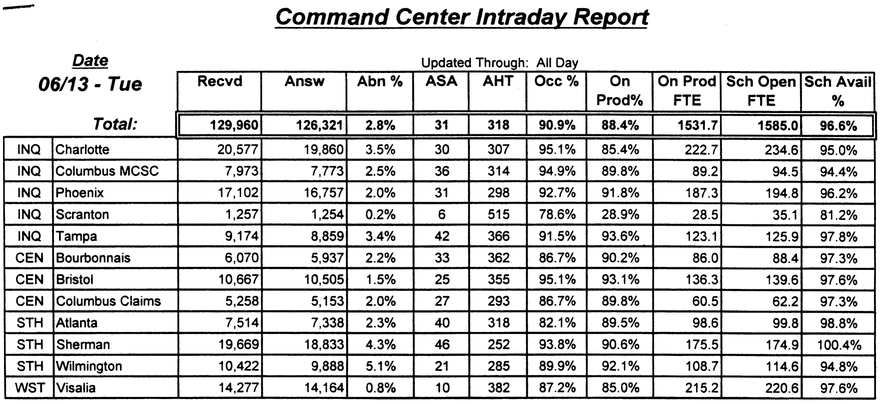
where $R=\frac{\lambda}{\mu}$ is the *offered load* (λ=arrival rate, μ=service rate) and β represents *service grade*. The actual value of β depends on the particular model and performance criterion used, but the form of s is extremely robust and accurate. As an example, for the $M/M/s$ queue analyzed in [11], β could be taken a positive function of the ratio between hourly staffing and delay costs, Δ is called the safety staffing. It is shown in [11] that the square-root principle is essentially asymptotically optimal for large heavily-loaded call centers ($\lambda\uparrow\infty$, $s\uparrow\infty$), and it prescribes operation in the rationalized (Halfin-Whitt) regime.

The square-root principle is applicable beyond $M/M/s$ (Erlang C). [24] verify it for the $M/M/s$ model with abandonment (Subsection 3.2) - here β can take also negative values, since abandonment guarantee stability at all staffing levels; for time-varying models, as in [31], β varies with time; and [12] uses it for skill-based routing. Finally, [43] supports the principle for the $M/G/s$ queue, given service times that are square integrable. (Extensions to heavy-tailed service times would plausibly give rise to safety staffing with power of R other than half.)

In all the extensions of [11], only the *form* $s=R+\beta\sqrt{R}$ was verified, theoretically or experimentally, but the determination of the exact value of β, based on economic considerations, is still an important open research problem. The square-root principle embodies another operational principle of utmost importance for call centers - economies of scale (EOS) - which we turn to.

Operational regimes and economies of scale Consider a typical situation that we encountered at a large U.S. mail-catalogue retailer. At the peak period of 10:00-11:00 a number of 765 customers customers called; service time is about 3.75 minutes on average with an after-call-work of 30 seconds and auxiliary work to the order of 5% of the time; ASA is about 1 seconds and only 1 call abandoned. But there were about 95 agents handling calls, resulting in about 65% utilization - clearly a *quality-driven* operation.

At the other extreme there are *efficiency-driven* call centers: with a similar offered work as above, ASA could reach many minutes and agents are utilized very close to 100% of their time.



Command Center Intraday Report

Date

06/13 - Tue

Updated Through: All Day

		Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
	Total:	**129,960**	**126,321**	**2.8%**	**31**	**318**	**90.9%**	**88.4%**	**1531.7**	**1585.0**	**96.6%**
INQ	Charlotte	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
INQ	Columbus MCSC	7,973	7,773	2.5%	36	314	94.9%	89.8%	89.2	94.5	94.4%
INQ	Phoenix	17,102	16,757	2.0%	31	298	92.7%	91.8%	187.3	194.8	96.2%
INQ	Scranton	1,257	1,254	0.2%	6	515	78.6%	28.9%	28.5	35.1	81.2%
INQ	Tampa	9,174	8,859	3.4%	42	366	91.5%	93.6%	123.1	125.9	97.8%
CEN	Bourbonnais	6,070	5,937	2.2%	33	362	86.7%	90.2%	86.0	88.4	97.3%
CEN	Bristol	10,667	10,505	1.5%	25	355	95.1%	93.1%	136.3	139.6	97.6%
CEN	Columbus Claims	5,258	5,153	2.0%	27	293	86.7%	89.8%	60.5	62.2	97.3%
STH	Atlanta	7,514	7,338	2.3%	40	318	82.1%	89.5%	98.6	99.8	98.8%
STH	Sherman	19,669	18,833	4.3%	46	252	93.8%	90.6%	175.5	174.9	100.4%
STH	Wilmington	10,422	9,888	5.1%	21	285	89.9%	92.1%	108.7	114.6	94.8%
WST	Visalia	14,277	14,164	0.8%	10	382	87.2%	85.0%	215.2	220.6	97.6%

Figure 2. Performance of 12 call centers in the rationalized regime.

Within the quality-driven regime, almost all customers are served immediately upon calling. At the efficiency-driven regime, on the other hand, essentially all customers are delayed in queue. However, as explained in [11] and elaborated on momentarily, well-managed large call centers operate within a *rationalized* regime, where quality and efficiency are balanced in the face of scale economies. This is the case in Figure 2, summarizing the performance of 12 call centers, operated by a large U.S. health insurance company: one observes a *daily* average of 2.8% abandonment (out of those called), 31 second ASA, 318 seconds AHT (Average Handling Time, namely service duration), with 91% agents' utilization (and over 95% in a couple of the call centers). Only about 40% of the customers were delayed while the other 60% accessed an agent immediately without any delay.

The rationalized regime was first identified in practice by Sze [48], from which we loosely quote the following: "The problems faced in the Bell System operator service differ from queueing models in the literature in several ways: 1. Server team sizes during the day are large, often 100-300 operators. 2. The target occupancies are high, but are not in the heavy traffic range. Approximations are available for heavy and light traffic systems, but our region of interest falls between the two. Typically, 90-95% of the operators are occupied during busy periods, but because of the large number of servers, only about half of the customers are delayed." Theory that supports the rationalized regime was first developed by Halfin and Whitt [28]. Thus large call centers operate in a regime that seems to circumvent the traditional tradeoff between service-level and resource-efficiency - EOS is the enabler.

As a practical illustration of EOS, consider multiple geographically dispersed call centers. By interconnecting them properly (dynamic load balancing), performance can get close to that of a single virtual call center, thus exploiting fully the economies of scale. This is the case in Figure 2, the header of which reads “Command Center Intraday Report”: and indeed, load balancing is exercised from a single Command Center that overseas the 12 call centers represented in the table. An ACD that distributes calls to several call centers is often referred to as a network-ACD.

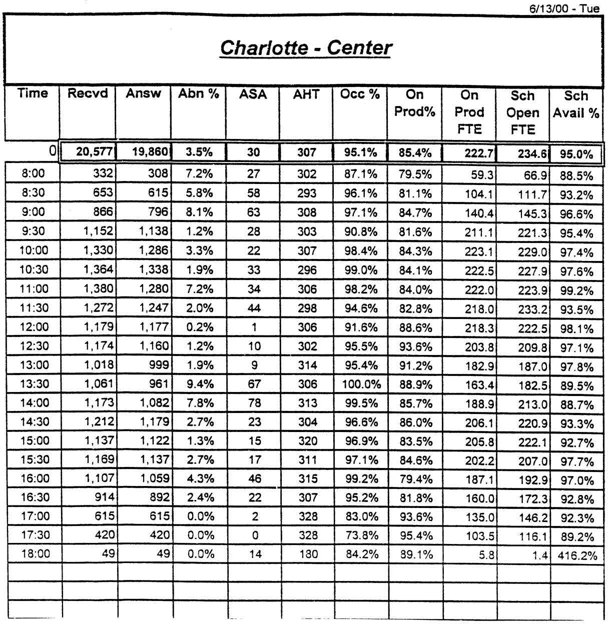
[46] analyzes the problem of setting routing probabilities, but more can be gained if routing is completely dynamic. [32] compares two basic strategies for a network-ACD: a centralized FIFO vs. a distributed strategy that routes an arriving call to the call center with least expected delay. Both strategies require information-exchange over the network. While FIFO is much more taxing, it could nevertheless be still inferior, given certain delays in switching calls between centers. This paper provides references to previous works on the subject, by the same group at AT&T.

3.2. Busy signals and abandonment

Each caller within a call center occupies a trunk-line. When all the lines are occupied, a calling customer gets a busy signal. Thus, a manager could eliminate *all* delays by dimensioning the number of lines to be equal to the number of agents. in which case $M/M/s/s$, or Erlang-B (“B” for Blocking) becomes the “right” model. But then there would typically be ample busy-signals. Moreover, prevailing practice goes in fact the other way: it is to dimension amlple lines so that a busy signal becomes a rare event. But then customers are forced into long delays. This is costly for the call center (think 1-800 costs) and possibly also for the customers - they might well prefer a busy-signal over an information-less delay, and hence they abandon the tele-queue before being served.

The busy-signal vs. delay vs. abandonment trade off has not yet been formally and fully analyzed, to the best of our knowledge. A simulation study of $M/M/s/B$ is presented in [20], where B stands for the overall number of lines ($B \geq s$); it is argued that only 10% lines in excess of agents provides good performance: more lines would give rise to too much waiting and fewer to too many busy signals. A more appropriate framework would be the $M/M/s/B+G$ queue, where $+G$ indicates arbitrarily distributed patience (following the notation and results of [7]). An analytically tractable model is the $M/M/s/B+M$, in which patience is assumed exponential. (For mathematical details see [44], pages 109–112, and [24].) Procedures for estimating the mean patience, as an input parameter to performance analysis, are given in [24,39]. Alternatively, mean patience could be used as a tuning parameter, where its value is determined to establish a fit between practice and theory - this will be the approach taken in the following example.

In heavy traffic, even a small fraction of busy-signals or abandonment could have a dramatic effect on performance, and hence must be accounted for. This will now be demonstrated via the $M/M/s+M$ model [41,7,24], which adds an abandonment feature



6/13/00 - Tue

Charlotte - Center

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
0	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	28	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1,286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	1,364	1,338	1.9%	33	296	99.0%	84.1%	222.5	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	1,174	1,160	1.2%	10	302	95.5%	93.6%	203.8	209.8	97.1%
13:00	1,018	999	1.9%	9	314	95.4%	91.2%	182.9	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	1,212	1,179	2.7%	23	304	96.6%	86.0%	206.1	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

Figure 3. Performance of a large call center in the rationalized regime.

to $M/M/s$ (Erlang C): specifically, one models customers' patience as exponentially distributed, independently of everything else; customers abandon if their patience expires before they reach an agent. We shall refer to the $M/M/s+M$ queue as Erlang A, "A" for Abandonment, and for the fact that this model interpolates between Erlang B and Erlang C.

A model for a call center with busy-signals should be $M/M/s/B+M$, to account for the existence of B lines. Performance analysis of the $M/M/s/B+M$ queue has been implemented at www.4callcenters.com. In this example, there were sufficiently many lines so that the busy signal phenomenon was negligible. We thus use Erlang A.

Consider Figure 3, which summarizes the daily operation of the Charlotte call center from Figure 2. Note the significant differences in performance over the busy half-hour periods while, on the other hand, the numbers of calling customers, as well as AHT and the number of agents working ("on production") do not seem to vary that significantly. Let us understand these performance differences. For example, during the period 10:30-11:00, the absence of only 5 agents (out of the 223 working) would likely result in almost doubling of both ASA and the fraction abandoning. We arrived at this projection by choosing the average of customers' patience (30 minutes) so that the predicted theoretical performance was close to the observed one. Interestingly and significantly, a model in which average patience is 30 minutes differs dramatically from a model which does not acknowledge abandonment ("infinite patience"): with our parameters, the latter would give rise to an unstable system (agents are required to be busy "more than 100%" of their time); stability could nevertheless be achieved by adding only 2 agents (225 all together), but in this case ASA would get close to 7 minutes - an order of magnitude error in predicting performance if one ignores abandonment (that is, if one uses Erlang C instead of Erlang A). We strongly recommend Erlang A as the standard to replace the prevalent Erlang C model.

[15] considers a call center with a finite number of lines, exponential patience and, prior to waiting, an IVR message of constant-duration. The model is thus a two-dimensional network, allowing for only approximations. Brandt & Brandt [14] solve the system with generally distributed patience (times to abandonment) and a finite number of lines. Also Brandt & Brandt [13] study a system with generally distributed patience and a secondary "call back" queue; again, this gives rise to approximations of a two-dimensional network.

[40] takes another perspective: they assume that rational customers compare their expected *remaining* waiting time with their subjective value of service. They provide evidence why rational callers should abandon at some time while being queued. Finally, [50] provides numerical evidence for the thesis of rational adaptive customers and present a new model for abandonment (simpler and more practical than that in [40]). For a discussion on service levels, including abandonment, we recommend [16].

Reality is even more complicated than described above, as demonstrated by the following reasoning. Decisions on agent staffing must take into account customer patience; the latter, in turn, is influenced by the waiting experience which, circularly, depends on staffing levels. An appropriate framework, therefore, is that of an equilibrium (Game Theory), arrived at through customer self-optimizing and learning. This is the perspective of [40] and [50], which constitutes merely a first step. In [40], abandonment arises as an equilibrium behavior of rational customers who optimaly compare their expected *remaining* waiting time with their subjective value of service. In [50], the model of [40] is simplified, which enables some support for adaptive behavior (learning) of customers.

Up to now we did not take into account the fact that callers that were blocked or that abandonned might try again at a later moment. This leads to retrial models (see [6,17,19]). Up to now retrial queues are little used in the context of call centers.

In [1], a model is considered where computer resources are assumed the bottlenecks, and hence they are explicitly modeled. Here all agents compete, in a processor sharing manner, for the same computer resource. This leads to certain counterintuitive phenomena: for example, performance levels could decrease as the number of agents increase. (In fact, [1] analyses a multi-skill environment.)

3.3. Performance over multiple intervals and overload

To make the translation to intra-day performance, and thus to inhomogeneous Poisson arrivals, (weighted) sums of interval performances are taken, where for each interval another call arrival rate is taken. [27] calls this the *pointwise stationary approximation.* An alternative idea would be to take the average arrival rate, and use this as input for a performance model. This can give extremely bad results, even if the occupancy is constant; see [26,27].

Standard modeling applications for call centers use stationary performance measures for each interval, say of 30 minutes duration. This works in general pretty well. But exceptions arise with abrupt significant changes in arrival rate, particularly when overload occurs during one or more intervals. Then a backlog is built up, and nonstationarity has to be accounted for. As already mentioned, such a behavior could arise from an external event, such as advertising a telephone number on TV, or when the call center opens in the middle of the day. Such abrupt overloads can be modeled with the help of fluid models, as in [37]. These results are extended in [38]. Unfortunately these fluid approximations work less well in underload situations, as has been argued in [3]. A numerical way to include nonstationary behavior is described in [22]. [31] proposes staffing guidelines, which were developed heuristically and gave rise to a time-varying square-root staffing principle.

3.4. Skill-based routing: on-line and off-line

The operational characteristics of multi-type/channel multi-skill contact centers could get very complicated [23]. Simply conceptualize a call center of say a large European company, which provides technical support in all major European languages for a broad product line. Nevertheless, and out of necessity, most call centers are multi-type multi-skill operations, and hence practice is here awaiting theoretical research for guidelines.

If each skill has dedicated agents, then of course the call center can be regarded as several independent single-skill call centers operating in parallel. But then one does not exploit the economies of scale, due to resource flexibility, of a large call center with multi-skill agents. At the other extreme, complete flexibility where all agents can do all tasks (for example, be able to support all products in all languages) is typically unrealistic. Thus a compromise must be struck where a subset of tasks, which we refer to as a *skill*, can be performed by a subgroup of agents - namely a *skill group.* Skills of different skill groups could overlap, which enables the benefits from economies of scale without the need to train all agents at all skills.

The operational challenges are then both off- and on-line. One should determine off-line the overall number of agents required of each skill, which are to be part of the company's permanent or temporary pool of agents; and out of these, how many and who should occupy a given shift. On-line, one should determine for an idling agent which caller to attend to first; and for an arriving call, who will be the agent to cater to it. In this section we survery on-line problems. The off-line issues are related to human resource management and are not discussed here.

Skill-based routing refers to the *on-line* strategy that matches callers and agents. It is nowadays part of any advanced ACD, often provided as a list of options that managers can choose from, but without any guidelines to accompany them. We now survey some related available research. For more information, readers are referred to the short literature survey in [23] and the OR and Simulation sections in [35].

[23] constitutes an introduction to skill-based routing and its operational complexities. Via simulation, it is demonstrated there that advantages can be considerable, already for simple scenarios. [42] provide a useful brief introduction to both theory and practice.

A common way of implementing skill-based routing is by specifying two selection rules: agent selection - how does an arriving call select an idle agent, if there is one; and call selection - how does an idle agent select a waiting call, if there is one. Here are some details. Agents are first divided into groups such that all agents within the group share the same skills. In general, several groups could have the same skill. The PABX/ACD contains, for each skill, an ordered list of agent groups containing that skill. An arriving call for a certain skill is then assigned to the first group in the list that has an agent available. When no agent with the right skill is available, then the call is assigned to the first agent with the skill that becomes available. If an available agent can handle each one of several waiting calls, then some priority rule is employed in order to determine which call to handle first. As far as we know, this common protocol has not been analyzed analytically.

If one leaves out the possibility that a call finds all agents occupied, then a flow of calls of a certain type from one agent group to the next group occurs only if all agents are occupied, i.e., it is overflow. These are notoriously hard to analyze, see [30], because the overflow process is not Poisson. The performance of this type of an overflow queueing network in the context of call centers is studied in [33].

It is also possible to program a PABX in such a way that a call is assigned to a group only if there is at least a certain threshold number of agents available for service. Thus agents are reserved idle for future high-priority calls while low-priority calls are presently waiting to be served. This becomes useful if a group has skills of varying importance, and it is advisable to reserve several agents free for the most important call types.

Although the above protocol is commonplace, it is certainly not optimal. E.g., it can occur that the last agent with skill A is occupied by a call of skill B, while there are multiple agents available with skills B and C. This effect cannot be avoided by changing the routing lists, due to the random behavior of the system. In fact, to reach optimal routing, one has to take the number of available agents in all groups into account. This way the routing becomes completely dynamic. The standard way to solve this type of problems is by Dynamic Programming. Unfortunately, it is impossible to apply standard Dynamic Programming to identify the optimal assignment, neither theoretically (the problem as of now seems too hard) nor practically, due to the so-called *curse of dimensionality* [8]: the number of possible configurations is exponential in the number of agent groups, making it numerically infeasible to apply standard algorithms from Markov decision theory. One way to overcome the problem's complexity is to consider simple structures and specific strategies. For example, [42] consider a two-channel system, where waiting customer are assigned an *aging factor*, proportional to their waiting time. Then customers with the largest aging factor is chosen for service. Alternatively, one could analyze provably-reasonable approximations, for example [12]. Both [42] and [12] consider the on-line routing problem as well as the of-line staffing problem - namely, how many agents are to be available for answering calls so as to maintain an acceptable grade of service. ([12] actually applies the square-root staffing principle.)

3.5. Call blending and multi-media

Different multi-media services require differing response times. Specifically, telephone services should be responded to within seconds or minutes and, once started, should not be interrupted; e.mail and fax, on the other hand, can be "stored" towards response within hours or days, and can definitely be preempted by telephone calls, and then resumed; chat services are somewhere in between. In [36] a mathematical asymptotic framework of Markovian Service Networks is developed, where multi-type customers are served according to preemptive-resume priority disciplines. The pitives of a Markovian service network are time-varying, abandoment and retrials are accomodated, and the asymptotics is in the rationalized (Halfin-Whitt) regime. The framework of [36] is thus applicable for performance analysis of large multi-media call centers - as indeed was done in [37,38]. Note however that the framework can not accommodate non-preemptive priority disciplines or finite buffers (busy-signals).

We now continue with models that include IVR and e.mail. Brandt and Brandt [13], already mentioned in the context of abandonment, propose a (birth-and-death) queueing model for a call center with impatient callers and an integrated IVR: callers that are patient enough, and which have been waiting online beyond a given threshold, are then transferred to ("stored in") an IVR-queue; the latter is served later, as soon as no customers are waiting online, and the number of idle agents exceeds another threshold. Armony and Maglaras [5] establish the asymptotic optimality in equilibrium of such a threshold strategy, when customers act rationally. By this we mean that customers who are not served immediately optimize among balking, abandoning, or opting for a return call (or a later e.mail) if they assess their anticipated delay as exceeding its worth. The equilibrium formulation is inspired (but differs from) [40,50]; the asymptotics is taken in the rationalized (Halfin-Whitt) regime.

If we mix traffic from multiple channels, then additional questions arise. Historically, these questions first arose in the context of mixing inbound and outbound traffic, but they are also applicable to multi-media traffic. The solution is called *call blending*, where agents are made to switch between inbound and outbound traffic, depending on the traffic loads of inbound traffic. A mathematical model for call blending is presented and solved in Bhulai & Koole [10].

Pure outbound Call centers are becoming more prevalent, mainly in surveys and tele-marketing. They use devices called *predictive dialers* that automatically call up customers, according to a prepared list. In order to reduce idleness of the most expensive call center resource, its agents, it often happens that the PABX calls the next customer on the list while, in fact, there are no agents available to take the call. Thus, the central problem is balancing between agent productivity (is there always a customer right away?) and customer dissatisfaction (no agent is idle while a customer picks up the phone), in a manner that is consistent with the company-specific relative importance of these two goals. For more information on predictive dialers, see Samuelson [47].

Acknowledgements G.K. would like to thank Sandjai Bhulai and Geert Jan Franx for their useful comments on the very first version of this paper, and an anonymous referee (of a different paper) for pointing out some sources of which he was not aware.

Some of the writing was done while A.M. was visiting Vrije Universiteit - the hospitality of G.K. and the institutional support are greatly appreciated. A.M. thanks Sergey Zeltyn for his direct and indirect contribution to the present project: Sergey helped in the preparation of the figures and tables, and he is the co-producer of the material from `ie.technion.ac.il/∼serveng` which was used here. Thanks are also due to Sergey and Anat Sakov for their approval of importing pieces of [39].

References

[1] O.Z. Akşin and P.T. Harker. Analysis of a processor shared loss system. *Management Science*, 47:324–336, 2001.

[2] O.Z. Akşin and P.T. Harker. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. Working paper, 2001.

[3] E. Altman, T. Jiménez, and G.M. Koole. On the comparison of queueing systems with their fluid limits. *Probability in the Engineering and Informational Sciences*, 15:165–178, 2001.

[4] B. Andrews and H. Parsons. Establishing telephone-agent staffing levels through economic optimization. *Interfaces*, 23(2):14–20, 1993.

[5] M. Armony and C. Maglaras. Customer contact centers with multiple service channels. Working paper, 2001.

[6] J.R. Artalejo. Accessible bibliography on retrial queues. *Mathematical and Computer Modelling*, 30:1–6, 1999.

[7] F. Baccelli and G. Hebuterne. On queues with impatient customers. In *Performance '81*, pages 159–179. North-Holland, 1981.

[8] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[9] L. Bennington, J. Commane, and P. Conn. Customer satisfaction and call centers: an Australian study. *International Journal of Service Industry Management*, 11:162–173, 2000.

[10] S. Bhulai and G.M. Koole. A queueing model for call blending in call centers. In *Proceedings of the 39th IEEE CDC*, pages 1421–1426. IEEE Control Society, 2000.

[11] S.C. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. Working paper, 2000.

[12] S.C. Borst and P. Seri. Robust algorithms for sharing agents with multiple skills. Working paper, 2000.

[13] A. Brandt and M. Brandt. On a two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability*, 1:191–210, 1999.

[14] A. Brandt and M. Brandt. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35:1–18, 1999.

[15] A. Brandt, M. Brandt, G. Spahl, and D. Weber. Modelling and optimization of call distribution systems. In V. Ramaswami and P.E. Wirth, editors, *Proceedings of the 15th International Teletraffic Conference*, pages 133–144. Elsevier Science, 1997.

[16] B. Cleveland and J. Mayben. *Call Center Management on Fast Forward.* Call Center Press, 1997.

[17] J.W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunications Review*, 18:49–100, 1957.

[18] A.K. Erlang. Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electroteknikeren*, 13:5–13, 1917. In Danish.

[19] G.I. Falin and J.G.C. Templeton. *Retrial Queues.* Chapman and Hall, 1997.

[20] M.A. Feinberg. Performance characteristics of automated call distribution systems. In *GLOBECOM '90*, pages 415–419. IEEE, 1990.

[21] R.A. Feinberg, I.-S. Kim, L. Hokama, K. de Ruyter, and C. Keen. Operational determinants of caller satisfaction in the call center. *International Journal of Service Industry Management*, 11:131–141, 2000.

[22] M.C. Fu, S.I. Marcus, and I-J. Wang. Monotone optimal policies for a transient queueing staffing problem. *Operations Research*, 48:327–331, 2000.

[23] O. Garnett and A. Mandelbaum. An introduction to skills-based routing and its operational complexities. Teaching note.

[24] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. Working paper.

[25] A. Gilmore and L. Moreland. Call centres: How can service quality be managed? *Irish Marketing Review*, 13:3–11, 2000.

[26] L. Green and P. Kolesar. Testing the validity of a queueing model of police patrol. *Management Science*, 37:84–97, 1989.

[27] L. Green and P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991.

[28] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–587, 1981.

[29] C.M. Harris, K.L. Hoffman, and P.B. Saunders. Modeling the irs telephone taxpayer information system. *Operations Research*, 35:504–523, 1987.

[30] A. Hordijk and A. Ridder. Stochastic inequalities for an overflow model. *Journal of Applied Probability*, 24:696–708, 1987.

[31] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42:1383–1394, 1996.

[32] Y. Kogan, Y. Levy, and R.A. Milito. Call routing to distributed queues: Is FIFO really better than MED? *Telecommunication Systems*, 7:299–312, 1997.

[33] G.M. Koole and J. Talim. Exponential approximation of multi-skill call centers architecture. In *Proceedings of QNETs 2000*, pages 23/1–10, 2000.

[34] B.W. Kort. Models and methods for evaluating customer acceptance of telephone connections. *IEEE*, pages 706–714, 1983.

[35] A. Mandelbaum. Call centers (centres): Research bibliography with abstracts. Electronically available as ie.technion.ac.il/∼serveng/References/ccbib.pdf, 2001.

[36] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998.

[37] A. Mandelbaum, W.A. Massey, M.I. Reiman, and R. Rider. Time varying multiserver queues with abandonments and retrials. In P. Key and D. Smith, editors, *Proceedings of the 16th International Teletraffic Conference*, 1999.

[38] A. Mandelbaum, W.A. Massey, M.I. Reiman, R. Rider, and A. Stolyar. Queue lengths and waiting times for multiserver queues with abandonment and retrials. Working paper, 2000.

[39] A. Mandelbaum, A. Sakov, and S. Zeltyn. Empirical analysis of a call center. Working paper, 2000.

[40] A. Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. *Queueing Systems*, 36:141–173, 2000.

[41] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4:189–208, 1953.

[42] M. Perry and A. Nilsson. Performance modeling of automatic call distributors: Assignable grade of service staffing. In *XIV International Switching Symposium*, pages 294–298, 1992.

[43] A.A. Puhalskii and M.I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32:564–595, 2000.

[44] J. Riordan. *Stochastic Service Systems*. Wiley, 1961.

[45] J.W. Roberts. Recent observations of subscriber behavior. In *Proceedings of the 9th International Teletraffic Conference*, 1979.

[46] L. Servi and S. Humair. Optimizing Bernoulli routing policies for balancing loads on call centers and minimizing transmission costs. *Journal of Optimization Theory and Applications*, 100:623–659, 1999.

[47] D.A. Sumuelson. Predictive dialing for outbound telephone call centers. *Interfaces*, 29(5):66–81, 1999.

[48] D.Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32:229–249, 1984.

[49] G. Tom, M. Burns, and Y. Zeng. Your life on hold: The effect of telephone waiting time on customer perception. *Journal of Direct Marketing*, 11:25–31, 1997.

[50] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and emperical support. working paper, 2000.