THE IMPACT OF INCREASED EMPLOYEE RETENTION UPON PERFORMANCE IN A CUSTOMER CONTACT CENTER

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027

December 1, 2004; Revision: October 10, 2005

Abstract

A mathematical model is developed to help analyze the benefit in contact-center performance obtained from increasing employee (agent) retention, by increasing agent job satisfaction. The contact-center "performance" may be restricted to a traditional productivity measure such as the number of calls answered per hour or it may include a broader measure of the quality of service, e.g., revenue earned per hour or the number of problems successfully resolved per hour. The analysis is based on an idealized model of a contact center, in which the number of employed agents is constant over time, assuming that a new agent is immediately hired to replace each departing agent. The agent employment periods are assumed to be independent and identically distributed random variables with a general *agent-retention probability distribution*, which depends upon management policy and actions. The *steady-state staff-experience distribution* is obtained from the agent-retention distribution by applying renewal theory. An increasing real-valued function specifies the average performance as a function of agent experience. Convenient closed-form expressions for the overall performance as a function of model elements are derived when either the agent-retention distribution or the performance function has exponential structure. Management actions may cause the agent-retention distribution to change. The model describes the consequences of such changes upon the long-run average staff experience and the long-run average performance.

Keywords: contact centers, call centers, retention, employee turnover, churn, agent job satisfaction, compensation, autonomy, stress, stochastic models, renewal theory, stochastic comparisons.

1. Introduction

It is widely recognized that contact-center performance is often hampered by low employee job satisfaction, as evidenced by high turnover, referred to as *churn* [11]. There is good reason to believe that churn can be reduced (retention can be increased) by increasing employee job satisfaction in various ways [6, 7, 13, 20, 21, 29, 30, 36].

The purpose of this paper is to develop a mathematical model that can provide insight into the way increased employee retention, achieved via increased employee job satisfaction, can improve performance. The employees we are thinking of are customer service representatives in contact centers, herein referred to as agents, but the analysis applies more broadly. For an overview of contact centers and various mathematical models that have been applied to them, see Gans et al. [16]. For a different mathematical model studying turnover, see Gans and Zhou [17]. For stochastic analysis of various behavioral aspects of queues, see Mandelbaum and Shimkin [23] and Zohar et al. [38].

We recognize that many of the issues surrounding agent job satisfaction and retention are not easily quantified. Nevertheless, we aim to quantify the performance benefits to be gained from increased agent retention. Moreover, we propose to take a relatively simple view, which allows us to focus carefully on a few critical issues. Our main thesis is that actions to increase agent job satisfaction (increasing autonomy or compensation, reducing stress, or by any other means) can benefit contact-center performance. Since agent job satisfaction is hard to measure, we view it through retention, which is directly observable (but subject to several possible definitions). We thus see increased agent job satisfaction improving performance in three steps: (i) increased agent job satisfaction increases agent retention; (ii) increased agent retention increases the staff experience, and (iii) increased staff experience increases performance.

We focus on productivity, and we focus on an easily measurable driver: *experience*, by which we mean simply time in service. (Clearly, there are other aspects of experience [26], but we do not consider them.) Increasing retention means that agents stay in service longer. When agents stay in service longer, the contact center tends to have a more experienced staff. We contend that an agent's performance, on average, should be an increasing function of the agent's experience. Staff experience directly influences performance, because performance typically improves dramatically through the initial start-up learning period. After that initial start-up learning period is over, we regard greater staff experience as an indication of greater

agent job satisfaction, which in turn should improve performance.

However, we recognize that, in general, the relation between turnover and performance is more complicated. There is evidence that performance can degrade when turnover is too low [2, 18]. There might be employee stagnation when the turnover is extremely low. Since the turnover is relatively high in many contact centers, we assume that the insufficient-turnover effect can be ignored. Hence, we assume that performance is simply an increasing function of experience (and thus a decreasing function of turnover).

We focus on the performance impact of increased retention. In doing so, we focus on only part of the story: When considering the many costs of employee turnover, it is natural to classify the costs, dividing them into two types: (i) transition costs, and (ii) productivity costs. Transition costs account for the per-agent cost of terminating the departing agent, recruiting and training a new agent, and disruption costs associated with the change, such as the cost of hiring a temporary employee, and the cost of managers coping with the change, such as the cost of performing exit interviews, the administrative costs of stopping benefit deductions and performing benefit enrollments, and so forth [8]. It has been estimated that the transition costs alone can be as much as $100\% - 200\%$ of an agent's annual compensation [8]. And yet we primarily focus on the productivity costs. We develop a mathematical model to describe the transition costs in Section 8. Clearly, a full analysis should include all costs and benefits of alternative policies to improve retention.

Our mathematical model will quantify how increased retention does indeed increase performance. As with all mathematical models, the value of its detailed quantitative conclusions depends on the appropriateness of the model assumptions and the model inputs. However, the mere process of modelling and analyzing can provide valuable insight.

It is common to discuss retention and churn in terms of a single number. For example, it may be said that the churn is 40% per year. Even though we take a rather narrow view of retention, we introduce a much more detailed model of agent retention: We assume that an agent's length of service is a random variable with a general *agent-retention probability distribution*. A probability distribution is used to account for individual differences among agents. Thus we characterize retention by a probability distribution, which is a function instead of a single number. We will also consider intermediate representations, in which the probability distribution is characterized by only a few parameters. Our approach is in the spirit of the long tradition of manpower planning models [5, 10, 19]. However, compared to that literature, our model is relatively elementary. Nevertheless, there are some novels steps here, in particular,

in the way we relate the agent retention (probability) distribution to the steady-state contact-center staff-experience distribution. In a reasonable contact-center scenario, we show how the agent-retention distribution determines the distribution of staff experience in the long run.

In practice, annual turnover is measured by dividing the number of agent terminations per year by the average staff size during the year. In the mathematical model, the annual turnover is the long-run rate of new hires per year divided by the (assumed) fixed number of agents, which is the reciprocal of the average length of employment for the agents (the mean of the agent-retention distribution). It is of course important that these different approaches are consistent, as we show in Section 8. As noted above, the rate of new hires will play a major role in estimating the important transition costs. The mean of the agent-retention distribution is a vital statistic, but we will show that the entire agent-retention distribution plays an important role for productivity. It is important that it is possible to estimate the agent-retention distribution from employment records. (For more discussion, see Section 7.) Our characterization of retention, even though somewhat elaborate, can be measured. We can measure what the agent-retention distribution has been and we can see how it changes.

We do less well in other aspects of the problem: First, here we do not consider specific management actions to increase agent sense of wellbeing or agent job satisfaction. (However, in a companion paper [31] we propose preference-based routing for that purpose.) Moreover, here we do not address how increased agent sense of wellbeing or increased agent job satisfaction increases agent retention, as measured by the agent-retention probability distribution. Those are important problems that remain to be investigated. We also do not determine how to measure agent performance. Instead, we simply assume that agent performance in fact can be measured and quantified. We could use a traditional productivity measure such as number of contacts handled per day, but we favor going beyond that to consider how well the agent helps the contact center meet its business objectives. Thus we think of performance depending on the revenue generated per day or the number of service requests successfully resolved per day. A good performance measure might be a weighted combination of several measures each focusing on a different aspect of performance. We are assuming that agent performance can be measured and quantified, but we do not address how to do it. That is a second problem that remains to be investigated.

Here is how the rest of the paper is organized: In Sections 2 and 3 we introduce our mathematical model. In Section 3 we introduce a probability model that allows us to relate the agent-retention probability distribution to the long-run distribution of staff experience.

We establish important properties of the long-run staff-experience distribution in Sections 4 and 6. We introduce tractable parametric models in Section 5. We discuss statistical fitting of the model elements in Section 7. We develop a mathematical model of the transition costs in Section 8. Finally, we draw conclusions in Section 9.

2. The Basic Mathematical Model

We assume that agent performance can be quantified and that quantification can be related to agent experience. In particular, we assume that there is a *performance function* r mapping experience (length of service) into average agent performance (appropriately specified), using r to suggest "revenue," "reward," "return," "rate of return" or "return on investment." We assume that the performance function is a nondecreasing function that approaches an asymptote (the maximum possible performance) as $t \to \infty$. Thus we can write

$$r(t) \equiv \rho R(t), \quad t \geq 0, \tag{2.1}$$

where $R(t) \to 1$ as $t \to \infty$. (We use $\equiv$ to denote equality by definition.) That makes R a probability *cumulative distribution function* (cdf); $R(t)$ is the proportion of the maximum possible expected performance, ρ, expected from an agent of experience t. We call R the *performance cdf* associated with the performance function r.

An example of a possible performance function is the *exponential performance function*

$$r(t) = \rho(1 - e^{-\lambda t}) \quad t \geq 0 \ , \tag{2.2}$$

which has the advantage of having only two parameters: ρ and λ; see p. 36 of [28]. The exponential performance function for $\rho = 10$ and $\lambda = 0.2$ is depicted in Figure 1. Note that the "scale" parameter ρ can be set arbitrarily, even if it is monetary, because we have yet to specify the units. We have chosen $\rho = 10$, thinking of a monetary reward rate measured in thousands of dollars per agent per month. Similarly, the parameter λ only acquires meaning when we specify the units. We have chosen $\lambda = 0.2$, thinking of time being measured in months. With that parameter value, significant progress occurs in the time scale of $1/\lambda$ which is 5 months.

We are interested in the overall long-run performance achieved by the contact center. To characterize that, we use the performance function r just defined and a *staff-experience cdf* F. For $x \geq 0$, $F(x)$ represents the long-run average proportion of agents with experience (term of employment) less than or equal to x months. We emphasize that $F(x)$ is intended to be a long-run average.

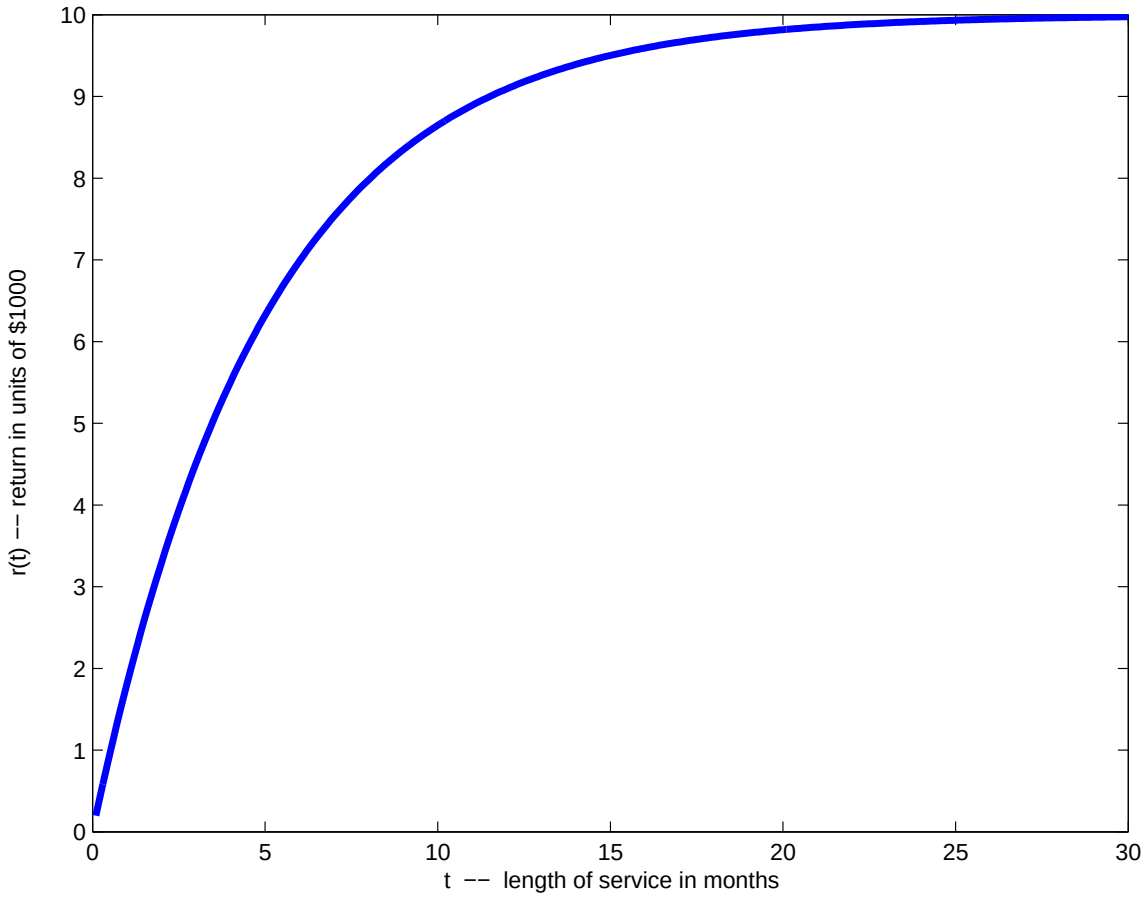

Figure 1: A possible performance function, mapping t, which measures agent experience in months (horizontal axis) into the selected performance, $r(t)$, which measures the return in thousands of dollars per month (vertical axis). Specifically, the exponential performance function in (2.2) is depicted with $\rho = 10$ and $\lambda = 0.2$.

At any given time t, the experience of the staff at that time is characterized by the *empirical staff-experience cdf*, denoted by F_t; $F_t(x)$ is the proportion of agents, at time t, who have been employed for a length of time less than or equal to x. At any given time, we can measure the empirical cdf F_t that describes the contact center at time t.

We can use the history of the empirical staff-experience cdf F_t to define the desired staff-experience cdf F: The staff-experience cdf F is the long-run average proportion of agents that have been in service for time less than or equal to x. For any x, the (long-run) staff-experience cdf $F(x)$ is the long-run average over time of the empirical staff-experience cdf $F_t(x)$; i.e., for all $x > 0$,

$$F(x) \equiv \lim_{T\to\infty} \frac{1}{T} \int_0^T F_t(x)\, dt \ . \tag{2.3}$$

In the next section, we will show that the staff-experience cdf F is well defined by (2.3) under our model assumptions. Moreover, unlike the empirical staff-experience cdf, the (long-run) staff-experience cdf F will have a *probability density function* (pdf), i.e., there is a function f such that

$$F(x) = \int_0^x f(u)\, du, \quad x \geq 0 \ . \tag{2.4}$$

We call f the *staff-experience pdf*. Unlike the empirical staff-experience cdf F_t, the staff experience cdf F and the associated staff-experience pdf f are deterministic functions.

We characterize the overall staff performance, denoted by $\mathbf{r}$, as the expected long-run average performance, i.e., the expected performance, r, weighted by the staff-experience pdf, f:

$$\mathbf{r} \equiv \mathbf{r}(r, f) \equiv \int_0^\infty r(t) f(t)\, dt. \tag{2.5}$$

There is another equivalent way to characterize this overall performance. Let A be a random variable with cdf F and pdf f. We think of A as the random experience (age) of a typical agent in the long-run. (We look at the system at an arbitrary time after the system has been operating for a long time and we pick an agent at random; A is the length of time that agent has been employed.) As before, $r(t)$ is the expected performance of an agent with experience (length of employment) t. Then (2.5) is equivalent to $\mathbf{r} = E[r(A)]$.

Paralleling the performance-function example above, an example of a staff-experience cdf F and associated staff-experience pdf f is the exponential distribution with mean $m_F = 1/\mu$, i.e., the exponential cdf and pdf

$$F(t) = 1 - e^{-\mu t} \quad \text{and} \quad f(t) = \mu e^{-\mu t}, \quad t \geq 0 \ , \tag{2.6}$$

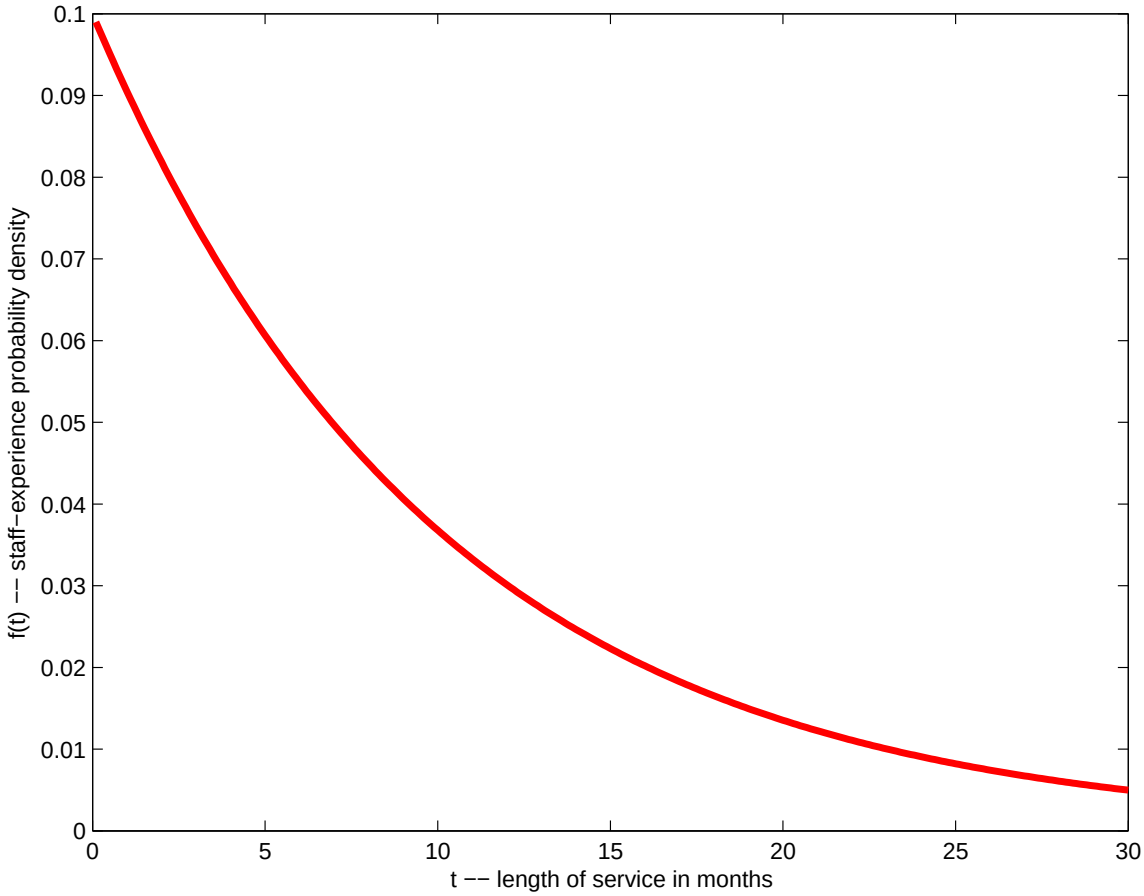

Figure 2: A possible staff-experience probability density function, mapping t, which measures agent experience in months (horizontal axis) into the probability density $f(t)$ of agents with experience t. Specifically, the exponential staff-experience pdf in (2.6) is depicted with $\mu = 0.1$ (mean $= 10$).

which has the single parameter μ. An exponential staff-experience pdf with mean 10 ($\mu = 0.1$) is depicted in Figure 2.

If we use *both* the exponential performance function r in (2.2) and the exponential staff-experience pdf f in (2.6), then we obtain a full model with three parameters: ρ, λ and μ. Then we can easily compute the overall performance. Then (2.5) becomes

$$\mathbf{r} = \int_0^\infty r(t)f(t)\,dt = \int_0^\infty \rho(1 - e^{-\lambda t})\mu e^{-\mu t}\,dt = \frac{\rho\lambda}{\lambda + \mu}\ . \tag{2.7}$$

From the simple formula in (2.7), we see how performance can be increased by increasing staff experience. Since the mean of the pdf f is $m(f) \equiv m_F = 1/\mu$, we *increase* staff experience when we *increase* the mean m_F. Formula (2.7) shows how the overall performance approaches the limit ρ as $m_F = 1/\mu$ increases.

For example, the overall long-run average performance $\mathbf{r}$ as a function of the mean $m(f) \equiv$

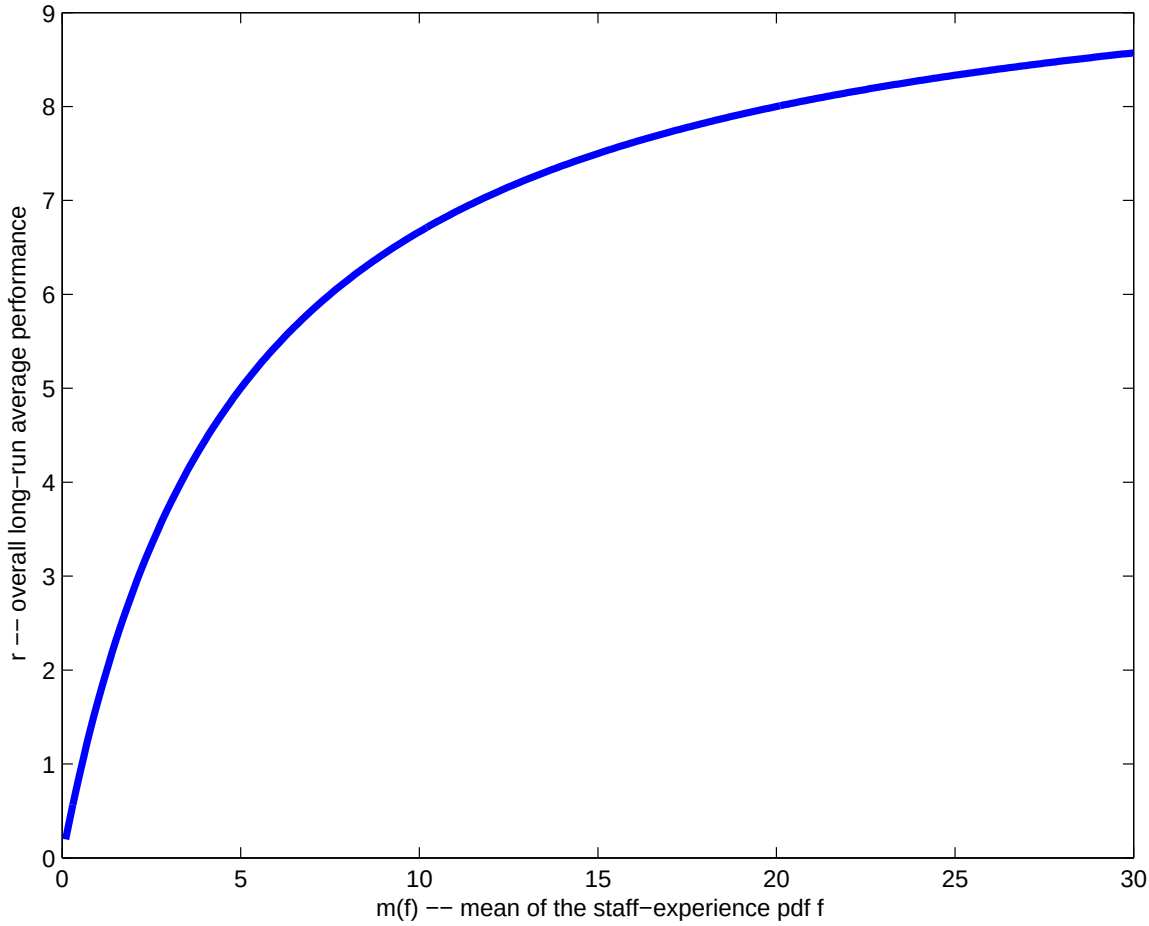

Figure 3: The overall long-run average performance $\mathbf{r}$ as a function of the expected long-run staff experience, $m \equiv m(f) \equiv m_F$, measured in months, for the totally exponential model with $\rho = 10$ and $\lambda = 0.2$ as in Figure 1 and $m(f) = 1/\mu$ for the exponential staff-experience pdf f.

m_F in the totally exponential model is plotted in Figure 3. Note that $\mathbf{r} \equiv \mathbf{r}(m_F)$ is

$$\mathbf{r}(m_F) = \frac{\rho\lambda m_F}{\lambda m_F + 1} \ . \tag{2.8}$$

For this example, $\rho = 10$ and $\lambda = 0.2$, so that $\mathbf{r}(m_F) = 2m_F/(0.2m_F + 1)$. From (2.8) or Figure 3, we see that the overall-performance function $\mathbf{r}(m_F)$ is increasing and concave: As we increase the mean staff-experience level m_F, the performance increases but the marginal gain decreases. Greater benefit from increasing the mean m_F occurs when m_F is lower.

It remains to relate the agent-retention pdf to the staff-experience pdf. That is the topic of the Section 3: In Section 3, we give an explicit formula for the staff-experience cdf F in terms of the agent-retention cdf G; see Theorems 3.1 and 3.2. It turns out that the staff-experience cdf is the renewal-process stationary-excess cdf associated with the agent-retention cdf. In Section 4 we show that relation enables us to deduce important properties of the staff-experience pdf f.

It turns out that, under our model assumptions, *the two cdf's coincide if and only if either of them is exponential*. Thus, for the exponential staff-experience pdf in (2.6), the agent-retention pdf must be exactly of the same form. So, if the agent-retention pdf is exponential, the analysis above applies. For example, Figure 3 is unchanged if the mean of the staff-experience pdf f on the horizontal axis is replaced by the mean of the agent-retention pdf.

On the other hand, *if the agent-retention pdf is not exponential, then the staff-experience pdf is neither exponential nor the same as the agent-retention pdf.* Nevertheless, as we just indicated, we derive an explicit formula for the staff-experience pdf in terms of the agent-retention pdf, so that a corresponding analysis can be carried out.

In Section 5 we introduce tractable parametric models, in which both the performance function r and the agent-retention pdf g (and thus the staff-experience pdf f) are more general than exponential, and yet the overall long-run average performance $\mathbf{r}$ can be represented as an explicit formula of the model parameters. We believe that a nice compromise between simplicity and flexibility is achieved when the two cdf's are each characterized by two parameters: the mean and the variance. We will show how the parametric models can be characterized in that way. Under regularity conditions, we show that the overall long-run average performance is an increasing function of the mean agent-retention time, when other parameters are appropriately held fixed.

In Section 6 we establish additional stochastic-comparison properties for the agent-retention cdf G and the staff-experience cdf F based on the relationship established in Section 3. We show that if the agent-retention cdf increases stochastically, in a sense to be made precise, then the staff-experience cdf increases stochastically as well, in a related way. However, we caution that the precise relationship requires careful definitions. We are thus able to give sufficient conditions for the overall performance to increase when the agent-retention cdf increases stochastically in an appropriate way.

3. The Agent-Retention Probability Model

We now connect the behavior of individual agents to the staff-experience cdf of the entire contact center. For that purpose, we make further assumptions. We consider an idealized model of a contact center, assuming that it contains a fixed number, n, of agents. We assume that a new agent is immediately hired to replace a departing agent whenever an agent departs.

Let $X_{i,k}$ be the length of time that the k^{th} agent in the i^{th} position is employed. We assume that the agent employment periods $X_{i,k}$ for $1 \le i \le n$ and $k \ge 1$, are independent

and identically distributed (IID) random variables distributed as a random variable X having a general cumulative distribution function (cdf) G – the *agent-retention cdf*, with probability density function (pdf) g – the *agent-retention pdf* – and finite k^{th} moment m_k for $k = 1, 2, 3$; i.e.,

$$G(t) \equiv P(X \le t) \equiv \int_0^t g(x)\,dx, \quad t \ge 0, \quad \text{and} \quad m_{G,k} \equiv E[X^k] \equiv \int_0^\infty t^k g(t)\,dt\ . \tag{3.1}$$

A possible agent-retention pdf g is depicted in Figure 4. It is a *gamma pdf*, i.e.,

$$g(t) = \frac{\mu(\mu t)^\nu e^{-\mu t}}{\Gamma(\nu)}\ , \quad t \ge 0\ , \tag{3.2}$$

where Γ is the Gamma function; see p. 37 of [28]. If ν is a positive integer, then $\Gamma(\nu) = (\nu-1)!$ A gamma distribution has two parameters: the *scale parameter* μ and the *shape parameter* ν. A gamma distribution has mean ν/μ, variance ν/μ^2 and thus *squared coefficient of variation* (SCV, variance divided by the square of the mean) $c_G^2 = 1/\nu$. The SCV is useful to measure variability independent of the mean. We can increase the mean without changing the SCV by decreasing the scale parameter μ; we can increase the variability, as measured by the SCV c_G^2, without changing the mean, by decreasing both ν and μ by the same amount. The particular gamma density shown in Figure 4 has mean $\nu/\mu = 10$, variance $\nu/\mu^2 = 50$ and SCV $c_G^2 = 1/\nu = 0.5$. With these parameter values, this gamma distribution coincides with an Erlang E_2 distribution [3, 37].

However, we caution that the gamma probability density function in Figure 4 may not have the correct shape. Studies have shown that the tendency to leave tends to decrease with time [14, 24]. Mathematically, that property can be expressed by saying that the agent-retention distribution should have *decreasing failure rate* (or hazard rate). If X is a random variable with cdf G and pdf g, the *failure rate* of X is

$$\lambda_X(t) \equiv \frac{g(t)}{G^c(t)}, \quad t \ge 0\ , \tag{3.3}$$

where $G^c(t) \equiv 1 - G(t)$ is the complementary cdf (ccdf) associated with the cdf G. Note that the hazard rate is the conditional intensity of a termination at t, given that the current length of service is t; see Chapter 4 of [4], p. 406 of [28] and Section 6 here, especially Definition 6.2, for more on DFR distributions. The empirical conclusions above mean that the failure-rate function $\lambda_X(t)$ should be a decreasing function of time t. A decreasing-failure-rate (DFR) distribution necessarily has a strictly decreasing pdf. Hence, agent-retention pdf's may look more like the the exponential pdf in Figure 2 (which has constant failure rate) than the gamma pdf in Figure 4. (Gamma distributions with $0 < \nu < 1$ are DFR, though.)

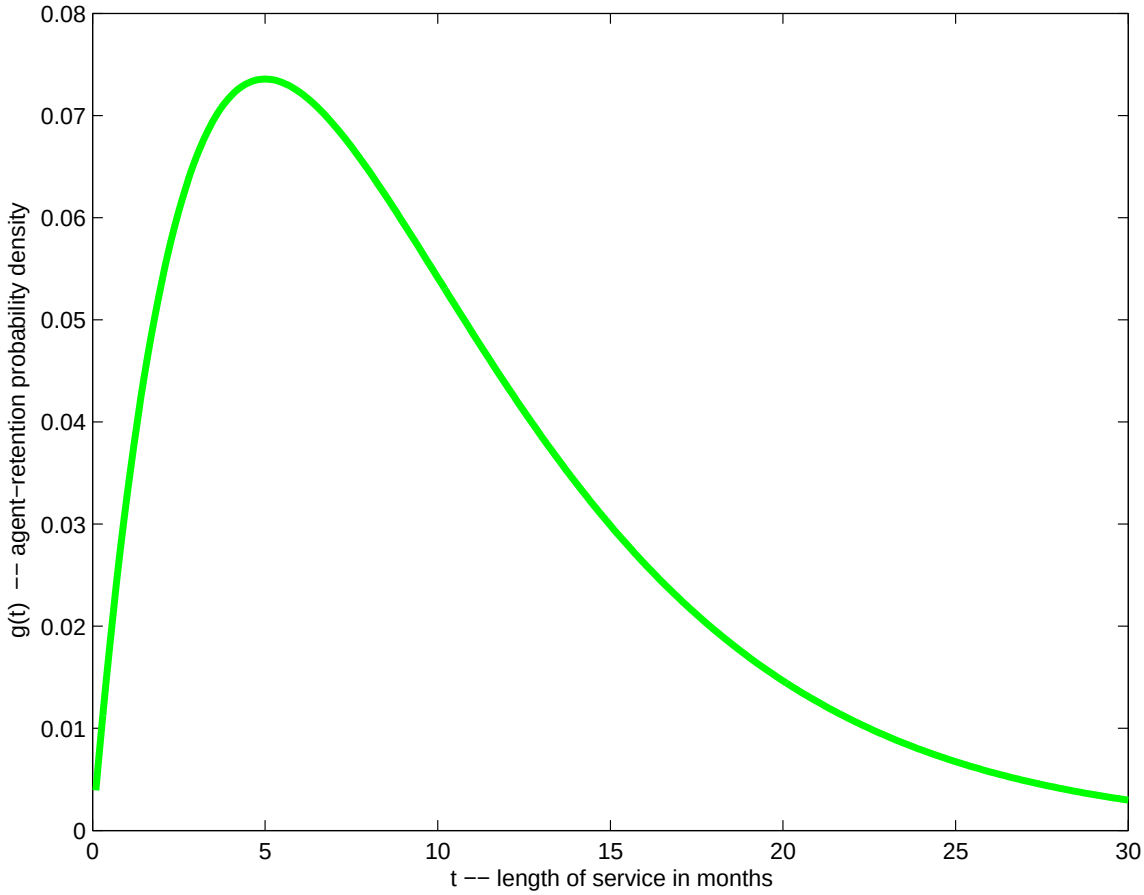

Figure 4: A possible agent-retention probability density function, mapping t, which measures time in months (horizontal axis) into the probability density $g(t)$ of an agent remaining employed for a length of time t. Specifically, a gamma agent-retention density in (2.6) is depicted with mean 10 and variance 50.

At any time t, the current experience of the center can be described by an n-dimensional random vector $\mathbf{A}(t) \equiv (A_1(t), \ldots, A_n(t))$, where $A_i(t)$ specifies the length of time that the agent in the i^{th} position has been employed, e.g., in months. We refer to $A_i(t)$ *age* or *experience* of the i^{th} agent at time t. The limiting steady-state distribution of the stochastic process $\mathbf{A} \equiv \{\mathbf{A}(t) : t \geq 0\}$ describes the experience of the contact-center staff in the long run.

Suppose that we hire n new agents at time $t = 0$. Then the n age processes $A_i \equiv \{A_i(t) : t \geq 0\}$ evolve as n IID age processes, also known as *backward recurrence-time processes*, associated with the *renewal process* with inter-renewal times distributed according to the agent-retention cdf G; see Section V.1 of Asmussen [3], Chapter 3 of Ross [27] and Chapter 7 of Ross [28] (especially Example 7.22 on p. 430). Thus, the stochastic process $\mathbf{A}$ is a Markov process and it has a proper limiting distribution as $t \to \infty$. The same limit also holds with a large class of alternative initial conditions. If we condition on particular initial ages, by assuming that $(A_1(0), \ldots, A_n(0)) = (y_1, \ldots, y_n)$, then the n age processes are again independent, but not identically distributed, age processes associated with *delayed renewal processes*, for which the same limit remains valid. See Asmussen [3] and Coffman et al. [12] for additional discussion and proofs. We formalize these established results in the following theorem. To state the result, let $\Rightarrow$ denote convergence in distribution for random vectors, e.g., see Chapters 3 and 11 in [35].

Theorem 3.1. *The vector-valued age stochastic process* $\mathbf{A}$ *is a Markov process. If, in addition to the conditions above,* $(A_1(0), \ldots, A_n(0)) = (y_1, \ldots, y_n)$ *for some vector* $(y_1, \ldots, y_n)$*, then*

$$\mathbf{A}(t) \Rightarrow \mathbf{Y} \quad \textit{as} \quad t \to \infty \ , \tag{3.4}$$

where, $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)$ *is a random vector with IID components (marginals), each distributed as a random variable* Y *having the classical stationary-excess distribution (or equilibrium residual-lifetime distribution)* G_e *associated with the cdf* G*, defined by*

$$G_e(t) \equiv \frac{1}{m_{G,1}} \int_0^t G^c(u)\, du \ , \tag{3.5}$$

where $G^c(t) \equiv 1 - G(t)$ *is the ccdf associated with* G *and* $m_{G,1}$ *is the mean of* G*; i.e., for all vectors of real numbers* $(x_1, \ldots, x_n)$*,*

$$P(Y_1 \leq x_1, Y_2 \leq x_2, \ldots, Y_n \leq x_n) = G_e(x_1)G_e(x_2)\cdots G_e(x_n) \ . \tag{3.6}$$

The cdf G_e *has* k^{th} *moment*

$$m_{G_e,k} \equiv E[Y^k] \equiv \int_0^\infty t^k g_e(t)\, dt = \frac{m_{G,k+1}}{(k+1)m_{G,1}}, \quad k \geq 1 \ . \tag{3.7}$$

It follows from the limit (3.4), the regenerative structure and the strong law of large numbers that the the time-average of the empirical staff-experience cdf F_t introduced in Section 1 converges with probability one, as desired in (2.3). Moreover, the limiting staff-experience cdf F is none other than the stationary-excess cdf G_e. To make the connection, note that

$$F_t(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[0,x]}(A_i(t)), \quad t \geq 0 \ , \tag{3.8}$$

where I_B is the indicator function of the set B; i.e., $I_B(x) = 1$ if $x \in B$, and $I_B(x) = 0$ otherwise. We formalize this important result as well.

Theorem 3.2. *Under the assumptions of this section, for all $x > 0$,*

$$P\left(\lim_{T\to\infty} \frac{1}{T}\int_0^T F_t(x)\,dt = G_e(x)\right) = 1 \tag{3.9}$$

for F_t in (3.8) *and G_e in* (3.5)*.*

Thus, for our model, the staff-experience cdf F introduced in Section 1 is well defined and equals G_e. Theorem 3.2 focuses on the limiting behavior of the *time-average* of the empirical staff-experience cdf F_t for any fixed number of agents, n. We can also focus on the limiting behavior of the steady-state empirical staff-experience cdf, as the number of agents, n, increases. Since the number of agents in a contact center is often large, it is interesting to consider that limit. The *steady-state empirical distribution* is

$$D_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n} I_{[0,x]}(Y_i), \quad x \geq 0 \ , \tag{3.10}$$

where Y_i is the steady-state limit of the age of the agent in the i^{th} position and I_B is again the indicator function of the set B. The classical strong law of large numbers (SLLN) and central limit theorem (CLT) imply the following result. Let $N(m, v)$ denote a normally distributed random variable with mean m and variance v. The SLLN uses convergence with probability one (w.p.1).

Theorem 3.3. *Under the conditions above, for each $x > 0$,*

$$D_n(x) \to G_e(x) \quad w.p.1 \quad as \quad n \to \infty \tag{3.11}$$

for D_n in (3.10) *and G_e in* (3.5)*, and*

$$\sqrt{n}[D_n(x) - G_e(x)] \Rightarrow N(0, \sigma^2(x)) \ , \tag{3.12}$$

where

$$\sigma^2(x) = G_e(x)(1 - G_e(x)) \ . \tag{3.13}$$

Theorem 3.2 says that the time-average of the empirical staff-experience cdf F_t approaches $F = G_e$ as $t \to \infty$, while the SLLN in (3.11) of Theorem 3.3 says that the steady-state empirical cdf D_n (after t has already become large) approaches $F = G_e$ as $n \to \infty$. Thus we have two different ways to justify focusing on the staff-experience cdf $F = G_e$.

We can also extend Theorem 3.3 to describe the limiting behavior *uniformly* in the argument x. We can apply the classical Glivenko-Cantelli theorem for that purpose.

Theorem 3.4. *Under the conditions above,*

$$P(\sup_{x \geq 0}\{|D_n(x) - G_e(x)|\} \to 0) = 1 \tag{3.14}$$

for D_n in (3.10) *and G_e in* (3.5).

There are corresponding stochastic generalizations paralleling the CLT in (3.12) related to the Kolmogorov-Smirnov statistic; e.g., see Section 2.2 of [35].

In summary, this section has presented a model, and analysis of that model based on renewal theory, showing that the staff-experience cdf F associated with an agent-retention cdf G should be the stationary-excess cdf G_e associated with the agent-retention cdf G, defined in (3.5).

4. The Staff-Experience PDF

In the previous section we saw that $F = G_e$. That enables us to deduce several important properties of the staff-experience cdf F. In particular, it enables us to deduce that the cdf F has a monotone (nonincreasing) pdf f.

Corollary 4.1. *Under the assumptions of Section* 3 (*even if G did not have a pdf*), *the staff-experience cdf F has a pdf f, i.e.,*

$$F(t) = \int_0^t f(u)\, du \ , \quad t \geq 0 \ , \tag{4.1}$$

where

$$f(t) = g_e(t) = (1/m_{G,1})G^c(t), \quad t \geq 0 \ . \tag{4.2}$$

Since the agent-retention cdf G has a pdf, the staff-experience pdf f is a continuous and nonincreasing function. Moreover, if $G^c(t) = 0$ for some t, then also $F^c(t) = 0$. On the other hand, if $G^c(t) > 0$ for some t, then $f(x) > 0$ for all x, $0 \leq x \leq t$.

Corollary 4.1 implies that the staff-experience pdf $f = g_e$ will be monotone nonincreasing, even though the agent-retention pdf g may fail to be monotone, as in Figure 4.

Since we have assumed that the agent-retention cdf G has a pdf g, we can describe the derivative of the staff-experience pdf f.

Corollary 4.2. *If the agent-retention cdf G has a pdf g, as assumed above, then the staff-experience pdf f is differentiable. The pdf f is convex if and only if the pdf g is nonincreasing.*

5. Parametric Models

We have given a general expression for the overall long-run average performance $\mathbf{r} \equiv \mathbf{r}(r, f)$ as a function of the performance function r and the staff-experience pdf f in (2.5). Since we are using the model in Section 3, we can replace f by $g_e = (1/m_1)G^c$. Hence, we can rewrite formula (2.5) as

$$\begin{aligned} \mathbf{r} &= \rho \int_0^\infty R(t) f(t)\, dt = \rho \int_0^\infty R(t) g_e(t)\, dt = \frac{\rho}{m_G} \int_0^\infty R(t) G^c(t)\, dt \\ &= \frac{\rho}{m_G} \left[\int_0^\infty G^c(t)\, dt - \int_0^\infty R^c(t)\, dt + \int_0^\infty R^c(t) G(t)\, dt \right] \\ &= \frac{\rho}{m_G} \left[m_G - m_R + \int_0^\infty R^c(t) G(t)\, dt \right] . \end{aligned} \tag{5.1}$$

Below we will use both the final expression and the last expression on the first line.

We have also given a closed-form expression for $\mathbf{r}$ in (2.7) for the case in which both r and f (and thus g) are both exponential functions in (2.7). In this section we give explicit formulas for more general parametric models. We especially want to characterize the two cdf's R and G by two parameters instead of one: the mean and SCV, instead of only the mean.

5.1. Hyperexponential Performance Functions

In this subsection we assume that the performance cdf R is a hyperexponential cdf. A hyperexponential (H_k) cdf is a mixture of k exponential cdf's. In particular, the H_k performance function is defined to be

$$r(t) = \rho R(t) , \tag{5.2}$$

where

$$R(t) = 1 - \sum_{i=1}^{k} p_i e^{-\lambda_i t}, \quad t \geq 0 , \tag{5.3}$$

with $p_i > 0$ and $\lambda_i > 0$ for each i, and $p_1 + \cdots + p_k = 1$.

For an H_k performance function, there are $2k+1$ parameters, one of which is the scale factor ρ and one of which is determined by the sum of the probabilities being equal to 1. Of course, the case $k=1$ yields the simple exponential performance function in (2.2). Since we have represented r in terms of the cdf R, we can use probabilistic methods. For example, a large class of cdf's (all completely monotone cdf's) can be represented as (not necessarily finite) mixtures of exponentials. Thus these cdf's can be approximated arbitrarily well by H_k cdf's, and an algorithm for doing so has been given by Feldmann and Whitt [15].

In applications, it may be of interest to consider H_k performance functions with $k=2$, because there are fewer parameters than for larger k, thus making it easier to fit. Indeed, there is a long history of using H_2 distributions to approximate probability distributions that are more variable than the exponential distribution. To reduce the number of remaining parameters from 3 to 2, it is common to let the H_2 distribution have *balanced means* by assuming that

$$\frac{p_1}{\lambda_1} = \frac{p_2}{\lambda_2} \ . \tag{5.4}$$

Provided that the SCV satisfies $c_R^2 > 1$, we can specify the parameters of the H_2 distribution in terms of the mean m_R and SCV c_R^2; see p. 137 of [33]:

$$p_i = \left[1 \pm \sqrt{(c_R^2-1)/(c_R^2+1)}\right]/2 \quad \text{and} \quad \lambda_i = \frac{2p_i}{m_R} \ . \tag{5.5}$$

It is not difficult to check that the H_2 distribution with the parameters in (5.5) has mean m_R, SCV c_R^2 and balanced means, as in (5.4). When $c_R^2 = 1$, we obtain the exponential distribution.

It is also possible to have a more general *three-parameter H_2 distribution*, which is fit to the first three moments of R: $m_{R,1}$, $m_{R,2}$ and $m_{R,3}$, provided that

$$m_{R,2} > 2m_{R,1}^2 \quad \text{and} \quad m_{R,3} \geq \frac{1.5m_{R,2}^2}{m_{R,1}} \ ; \tag{5.6}$$

see p. 136 of [33] and p. 592 of [1]. To obtain the third parameter, we drop the balanced-means condition in (5.4).

We now give an explicit expression for the overall performance with an H_k performance function. For that purpose, let $\hat{h}$ be the *Laplace transform* of the real-valued function h of a positive real variable, defined by

$$\hat{h}(s) \equiv \int_0^\infty e^{-sx} h(x)\, dx \ , \tag{5.7}$$

where s is a complex variable with positive real part. We will consider Laplace transforms with real-variable arguments.

Theorem 5.1. *For an H_k performance function, as in* (5.2)–(5.3)*, the overall performance is*

$$\mathbf{r} = \rho \left(1 - \sum_{i=1}^{k} p_i \hat{g}_e(\lambda_i) \right) = \frac{\rho}{m_G} \left(m_G - m_R + \sum_{i=1}^{k} \frac{p_i \hat{g}(\lambda_i)}{\lambda_i} \right) \ , \tag{5.8}$$

where $m_G \equiv m_{G,1}$ is the mean of G and m_R is the mean of R, i.e.,

$$m_R \equiv m_{R,1} = \sum_{i=1}^{k} (p_i/\lambda_i) \ . \tag{5.9}$$

Proof. We combine (5.1), (5.2) and (5.3) with well known relationship among $\hat{g}_e$, $\hat{G}$ and $\hat{g}$:

$$\hat{G}(s) \equiv \int_0^\infty e^{-sx} G(x)\, dx = \frac{\hat{g}(s)}{s} \quad \text{and} \quad \hat{g}_e(s) \equiv \int_0^\infty e^{-sx} \frac{(1 - G(x))}{m_G}\, dx = \frac{(1 - \hat{g}(s))}{s m_G} \ . \quad \blacksquare \tag{5.10}$$

5.2. Hyperexponential Agent-Retention Distributions

In this subsection we let the performance cdf R be general, but let the agent-retention cdf G be a hyperexponential cdf. As indicated in Section 3, this seems to be consistent with empirical research, because all hyperexponential distributions are DFR.

In particular, paralleling (5.3), we assume that

$$G(t) = 1 - \sum_{j=1}^{l} q_j e^{-\mu_j t}, \quad t \geq 0 \ , \tag{5.11}$$

where $q_j > 0$ and $\mu_j > 0$ for each j, and $q_1 + \cdots + q_l = 1$.

That implies that the associated stationary-excess cdf G_e is also H_k. in particular,

$$G_e(t) = 1 - \frac{1}{m_G} \sum_{j=1}^{l} (q_j/\mu_j) e^{-\mu_j t}, \quad t \geq 0 \ , \tag{5.12}$$

where $m_G \equiv m_{G,1}$ is the mean of G; here

$$m_G = \sum_{j=1}^{l} (q_j/\mu_j). \tag{5.13}$$

Consequently, closely paralleling the previous subsection, we see that g_e has exponential structure, so that we have the following result.

Theorem 5.2. *For a hyperexponential agent-retention cdf G, as in* (5.11)*, the overall performance is*

$$\mathbf{r} = \frac{\rho}{m_G} \int_0^\infty R(t) \sum_{j=1}^{l} q_j e^{-\mu_j t}\, dt = \frac{\rho}{m_G} \sum_{j=1}^{l} q_j \hat{R}(\mu_j) \ , \tag{5.14}$$

where $\hat{R}(s)$ is the Laplace transform of the cdf R.

5.3. The HHRP Model

In this section we combine the two hyperexponential assumptions made in the previous two subsections: We assume that *both* the performance cdf R and the agent-ratention cdf G are hyperexponential distributions; i.e., we assume that (5.2), (5.3) and (5.11) all hold. We call the resulting model the *hyperexponential-hyperexponential retention-performance* (HHRP) *model*.

From (5.8), we get

$$\mathbf{r} = \frac{\rho}{m_G}\left(m_G - m_R + \sum_{i=1}^{k}\frac{p_i}{\lambda_i}\sum_{j=1}^{l}\frac{q_j\mu_j}{\mu_j+\lambda_i}\right) \ . \tag{5.15}$$

On the other hand, from (5.14), we get

$$\mathbf{r} = \frac{\rho}{m_G}\sum_{j=1}^{l}q_j\sum_{i=1}^{k}\frac{p_i\lambda_i}{\mu_j(\lambda_i+\mu_j)} = \frac{\rho}{m_G}\sum_{j=1}^{l}\frac{q_j}{\mu_j}\sum_{i=1}^{k}\frac{p_i\lambda_i}{(\lambda_i+\mu_j)} \ . \tag{5.16}$$

Algebraic manipulations show that these two representations are equivalent.

When we use the three-parameter H_2 distributions, we obtain an overall model with 7 parameters: ρ, λ_1, λ_1, p_1, μ_1, μ_2 and q_1. (We know $p_2 = 1-p_1$ and $q_2 = 1-q_1$.) When we use the two-parameter H_2 fit with balanced means, based on (5.4) and (5.5), we obtain a model with five parameters: ρ, m_G, m_R, c_G^2 and c_R^2, where c^2 is the SCV. The H_2 distributions are always more variable than an exponential distribution, so that we have the constraint $c^2 \geq 1$. The H_2 distribution reduces to a single exponential distribution when $c^2 = 1$.

5.4. The GHRP Model

In this section and the next we develop alternative models in which one of the two hyperexponential distributions in the HHRP model is replaced by a gamma distribution. We are motivated to consider the gamma distribution, because it allows all possible positive SCV's; we can have $0 < c^2 < 1$ in addition to $c^2 \geq 1$. Hence with the three distributions - gamma, exponential and hyperexponential - we provide distributions that can be fit to all possible positive means and SCV's. We could do that with just the gamma distribution, but the hyperexponential distribution is easier to work with.

In this section we consider a gamma agent-retention distribution, and thus obtain the *gamma-hyperexponential retention-performance* (GHRP) *model*. In addition to the H_k performance function introduced in Subsection 5.1, a gamma pdf is used for the agent-retention pdf g. The gamma pdf is given in (3.2). The key additional fact is that the gamma pdf has a

convenient explicit Laplace transform. In particular, if g has a gamma pdf with parameters μ and ν, as in (3.2), then

$$\hat{g}(s) = \left(\frac{\mu}{\mu + s}\right)^{\nu} . \tag{5.17}$$

Hence, for the GHRP model,

$$\mathbf{r} = \frac{\rho}{m_G}\left(m_G - m_R + \sum_{i=1}^{k} \frac{p_i \mu^{\nu}}{\lambda_i(\mu + \lambda_i)^{\nu}}\right) . \tag{5.18}$$

For the gamma agent-retention pdf g in (3.2), we increase the mean $m_G \equiv m_{G,1} = \nu/\mu$, while holding the SCV $c_G^2 = 1/\nu$ fixed, if we decrease μ. It is thus natural to look at $\mathbf{r} \equiv \mathbf{r}(m_G)$ as a function of the mean m_G alone, with the understanding that we increase m_G by decreasing μ, while holding the shape parameter ν fixed.

We illustrate by displaying $\mathbf{r}(m_G)$ as a function of m_G, with the shape parameter ν held fixed, for a concrete example in Figure 5 below: We use an H_2 performance function r with parameters: $\rho = 10$, mean $m_R = 5.0$, SCV $c_R^2 = 2.0$ and balanced means, as in (5.4); we apply (5.5) to get the H_2 parameters $p_1, p_2, \lambda_1, \lambda_2$. We use a gamma agent-retention distribution with SCV $c_G^2 = 1/\nu = 0.5$, which corresponds to an Erlang (E_2) distribution. We let the mean of G, m_G, vary from 0 to 60 months, and see what happens to the long-run average performance $\mathbf{r}(m_G)$; i.e., we are plotting (5.18), letting m_G vary (by decreasing μ). We see that the overall long-run average performance increases towards its maximum value $\rho = 10$ as m_G increases. Moreover, we see that the function $\mathbf{r}(m_G)$ is concave, showing that the marginal gain decreases as m_G increases.

The concrete formulas we have derived let us study the impact of the different parameters on the overall long-run average performance. We illustrate by repeating the GHRP-model example above, considering three different SCV's for the gamma agent-retention cdf G: 0.25, 1.00 and 4.00. In Figure 6, we see more rapid convergence to the maximum possible long-run average performance ($\rho = 10$) with greater variability (higher c_G^2).

5.5. The HGRP Model

In this subsection and we consider the *hyperexponential-gamma retention-performance* (HGRP) *model*, obtained by using a hyperexponential agent-retention pdf g and a gamma performance cdf R; i.e., we now switch the roles of the two distributions used in the previous subsection.

Thus, let the hyperexponential agent-retention cdf G be as in (5.11). Now that the performance function is gamma, let the performance cdf R have Laplace transform

$$\hat{R}(s) = \frac{1}{s}\left(\frac{\lambda}{\lambda + s}\right)^{\nu} . \tag{5.19}$$

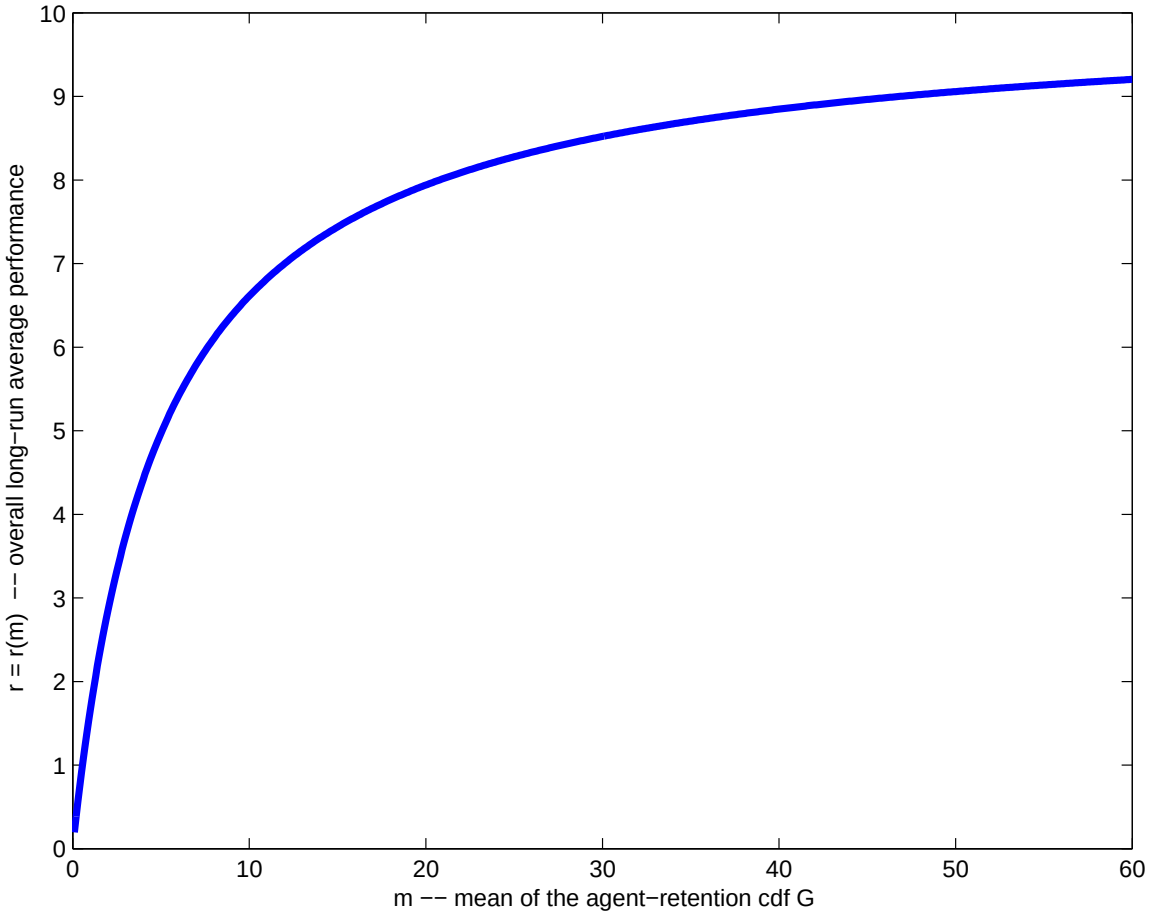

Figure 5: The overall long-run average performance $\mathbf{r} \equiv \mathbf{r}(m)$ in the GHRP model as a function of $m \equiv m_G$, the mean of the agent-retention cdf G, with the gamma agent-retention pdf in (3.2) having SCV $c_G^2 = 0.5$ (corresponding to an Erlang E_2 distribution) and H_2 performance function in (5.2)–(5.3) with $k = 2$, $\rho = 10$, mean $m_R = 5.0$, SCV $c_R^2 = 2.0$ and balanced means, as in (5.4).

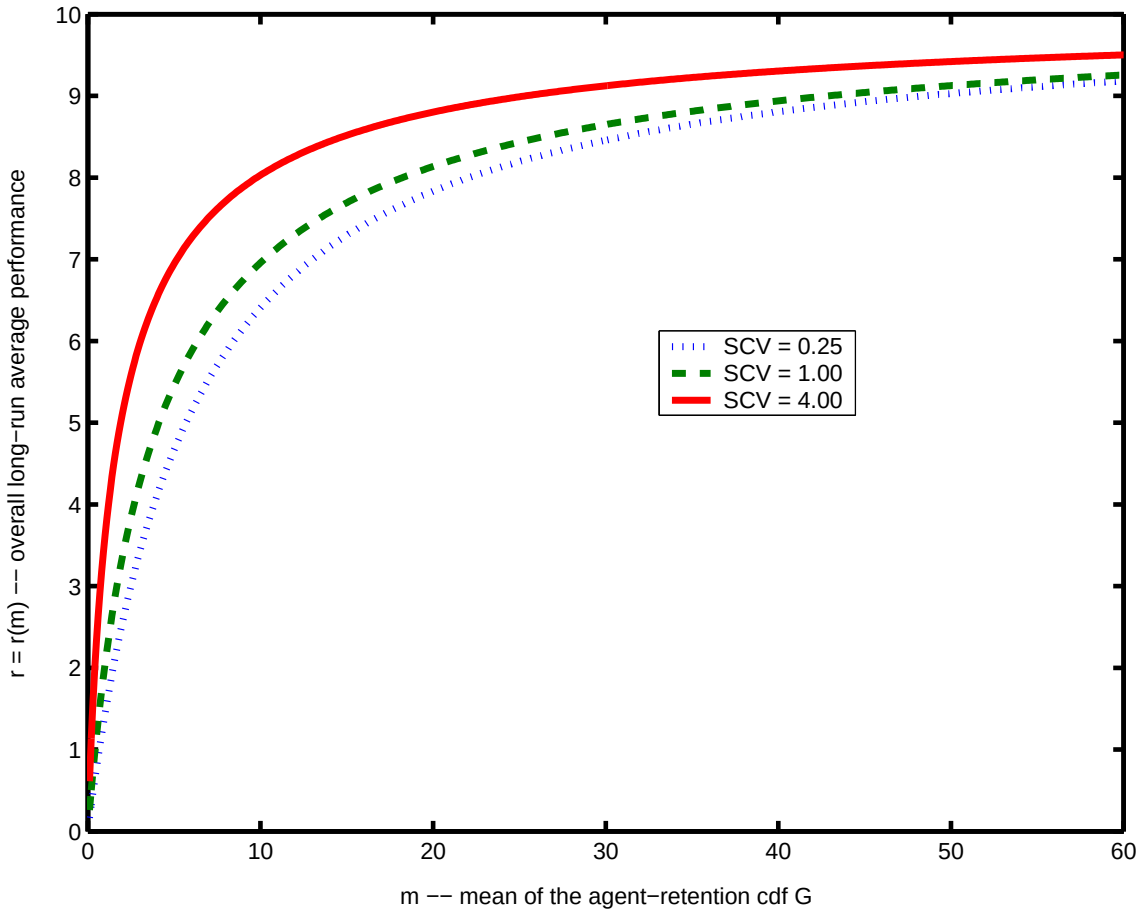

Figure 6: The overall long-run average performance $\mathbf{r} \equiv \mathbf{r}(m)$ in the GHRP model as a function of $m \equiv m_G$, the mean of the agent-retention cdf G, with the gamma agent-retention pdf in (3.2) for three different agent-retention pdf's: having SCV $c_G^2 = 0.25$, $c_G^2 = 1.00$ and $c_G^2 = 4.00$. The H_2 performance function is the same as in Figure 5.

We thus can apply (5.14) to get the final form

$$\mathbf{r} = \frac{\rho}{m_G} \sum_{j=1}^{l} \frac{q_j}{\mu_j} \left(\frac{\lambda}{\lambda + \mu_j} \right)^{\nu} . \tag{5.20}$$

6. Stochastic Comparisons

Our main goal in this section is to show that the overall long-run average performance $\mathbf{r}$ increases if the agent-retention cdf G increases stochastically in an appropriate way. For this purpose, we review basic stochastic-comparison concepts; see Chapter 9 of Ross [27] and Müller and Stoyan [25]. As in previous sections, we are drawing on established results.

We write $X_1 \leq_{SO} X_2$ or $G_1 \leq_{SO} G_2$ if X_i is a real-valued random variable with cdf G_i for $i = 1, 2$ and the probability distribution of X_1 (characterized by G_1) is stochastically less than or equal to the probability distribution of X_2 (characterized by G_2) in a sense denoted by $\leq_{SO}$, which remains to be defined. We also write $X_1 \geq_{SO} X_2$ or $G_1 \geq_{SO} G_2$ if $X_2 \leq_{SO} X_1$ or $G_2 \leq_{SO} G_1$.

To state the definitions we will use, for each i, let $G_i(t) \equiv P(X_i \leq t)$ be the cdf of X_i; let g_i be the pdf of the cdf G_i, assumed to be well defined and positive on the positive half line $[0, \infty)$; let $G_i^c(t) = 1 - G_i(t)$ be the associated ccdf; let $G_{i,e}$ be the associated stationary-excess cdf, defined as in (3.5); let $X_{i,t}$ be a random variable with (conditional) cdf $G_{i,t}(x) \equiv P(X_i \leq t + x | X_i > t)$ for $x > 0$ and $t > 0$; and let λ_i be the *hazard-rate function* (or failure-rate function), defined as in (3.3) by

$$\lambda_i(t) \equiv \frac{g_i(t)}{G_i^c(t)}, \quad t \geq 0 , \tag{6.1}$$

for all t such that $G_i^c(t) > 0$. We summarize the definitions of several stochastic orderings in the following definition. For some of the definitions, we state two equivalent characterizations.

Definition 6.1. *Five different notions of stochastic order are:*

(*ordinary*) *stochastic order:*

(*a*) $X_1 \leq_{ST} X_2$ *if* $G_1^c(t) \leq G_2^c(t)$ *for all* t ;

(*a′*) $X_1 \leq_{ST} X_2$ *if* $E[f(X_1)] \leq E[f(X_2)])$ *for all nondecreasing real-valued* f ;

increasing-convex (*stochastic*) *order:*

(*b*) $X_1 \leq_{IC} X_2$ *if* $\int_x^\infty G_1^c(t)\, dt \leq \int_x^\infty G_2^c(t)\, dt$ *for all* $x \geq 0$;

(b') $X_1 \leq_{IC} X_2$ *if* $E[f(X_1)] \leq E[f(X_2)])$ *for all nondecreasing convex real-valued* f ;

convex (stochastic) order (or variability order):

(c) $X_1 \leq_C X_2$ *if* $X_1 \leq_{IC} X_2$ *and* $E[X_1] = E[X_2]$;

(c') $X_1 \leq_{IC} X_2$ *if* $E[f(X_1)] \leq E[f(X_2)])$ *for all convex real-valued* f ;

hazard-rate (stochastic) order:

(d) $X_1 \leq_H X_2$ *if* $\lambda_1(t) \geq \lambda_1(t)$ *for all* t ;

likelihood-ratio (stochastic) order:

(e) $X_1 \leq_{LR} X_2$ *if* $\frac{g_1(t)}{g_1(s)} \leq \frac{g_2(t)}{g_2(s)}$ *for all* $0 \leq s < t$.

We now summarize established relations among these different notions of stochastic order. We write $\leq_{O1} \rightarrow \leq_{O2}$ if $X_1 \leq_{O1} X_2$ implies that $X_1 \leq_{O2} X_2$. The following represents all possible implications among these stochastic-order relations:

$$\leq_{LR} \rightarrow \leq H \rightarrow \leq_{ST} \rightarrow \leq_{IC} \quad \text{and} \quad \leq_C \rightarrow \leq_{IC} \ , \tag{6.2}$$

with the understanding that implications extend by transitivity.

We now define properties of individual probability distributions.

Definition 6.2. *The following are definitions of properties of the distribution of a random variable X with cdf G, pdf g, hazard-rate function λ and conditional residual-lifetime cdt $G_t(x) \equiv P(X_t \leq x) \equiv P(X \leq x + t | X > t)$ for $t \geq 0$:*

(a) *G has* ***increasing failure rate*** *(is IFR) if $\lambda(t)$ is a nondecreasing function of t;*

(a') *G has* ***decreasing failure rate*** *(is DFR) if $\lambda(t)$ is a nonincreasing function of t;*

(b) *G has a* ***new-better-than-used*** *(NBU) distribution if $G_t \leq_{ST} G$ for all t;*

(b') *G has a* ***new-worse-than-used*** *(NWU) distribution if $G_t \geq_{ST} G$ for all t;*

(c) *G has a* ***new-better-than-used-in-expectation*** *(NBUE) distribution if $E[X_t] \leq E[X]$ for all t;*

(c') *G has a* ***new-worse-than-used-in-expectation*** *(NWUE) distribution if $E[X_t] \geq E[X]$ for all t.*

We now summarize established relations among these different properties. We write $Prop_1 \rightarrow Prop_2$ if X (G) has property $Prop_2$ whenever it has property $Prop_1$. The following represents all possible implications among these properties:

$$IFR \rightarrow NBU \rightarrow NBUE \quad \text{and} \quad DFR \rightarrow NWU \rightarrow NWUE \ . \tag{6.3}$$

with the understanding that implications extend by transitivity.

Now we are ready to state established results about the relation between G and G_e. The following are conditions for stochastic comparisons between the cdf's G_e and G:

$$G_e \leq_{ST} (\geq_{ST}) G \quad \text{if and only if} \quad G \quad \text{is NBUE (NWUE)} \ ; \tag{6.4}$$

$$G_e \leq_{LR} (\geq_{LR}) G \quad \text{if and only if} \quad G \quad \text{is IFR (DFR)} \ ; \tag{6.5}$$

$$G = G_e \quad \text{if and only if} \quad G \quad \text{is exponential} \ . \tag{6.6}$$

For (6.4), see Problem 9.27 of [27].

Now we state established comparison results for the stationary-excess cdf's $G_{1,e}$ and $G_{2,e}$ associated with two different agent-retention cdf's G_1 and G_2. The following are conditions for stochastic comparisons between the cdf's $G_{1,e}$ and $G_{2,e}$:

$$G_{1,e} \leq_{LR} G_{2,e} \quad \text{if and only if} \quad G_1 \leq_H G_2 \ , \tag{6.7}$$

so that we have the implications

$$G_1 \leq_{LR} G_2 \ \rightarrow \ G_1 \leq_H G_2 \ \rightarrow \ G_{1,e} \leq_{LR} G_{2,e} \ \rightarrow \ G_{1,e} \leq_H G_{2,e} \ . \tag{6.8}$$

For (6.7), see Problem 9.18 of [27].

If $E[X_1] = E[X_2]$, then

$$G_{1,e} \leq_{ST} G_{2,e} \quad \text{if and only if} \quad G_1 \leq_{IC} G_2 \ , \tag{6.9}$$

so that

$$\text{if} \quad G_1 \leq_C G_2, \quad \text{then} \quad G_{1,e} \leq_{ST} G_{2,e} \ . \tag{6.10}$$

Finally, we have the following result about the way the overall performance depends upon the agent-retention cdf G.

Theorem 6.1. *Suppose that the performance function r is a nondecreasing real-valued function. If either*

$$G_1 \leq_H G_2 \quad \textit{or} \quad G_1 \leq_C G_2, \tag{6.11}$$

then

$$\mathbf{r}_1 = \int_0^\infty r_1(t) g_{1,e}(t)\, dt \leq \int_0^\infty r_2(t) g_{2,e}(t)\, dt = \mathbf{r}_2 \ . \tag{6.12}$$

Proof. The conclusion (6.12) holds if and only if $G_{1,e} \leq_{ST} G_{2,e}$ by Definition 6.1 (a'). However, that is implied by each of the conditions in (6.11), by virtue of (6.8) and (6.10), using Theorem **??**. ∎

Since a scalar multiple cX is exponential (H_k) whenever X is exponential (H_k), it is natural to consider the the distribution of cX as a function of c. We now give sufficient conditions for the distribution of cX to be increasing in c in the ordering $\leq_H$, which implies that the overall long-run average performance will increase when we multiply X by a constant $c > 1$, by virtue of Theorem 6.1.

Theorem 6.2. *Let X be a random variable with the agent-retention cdf G. If G has a failure-rate function λ_X satisfying*

$$\lambda_X(cx) \leq c\lambda_X(x) \quad \textit{for all} \quad x > 0 \quad \textit{and} \quad c > 1 \ , \tag{6.13}$$

which is implied by G being DFR, then the distribution of cX is increasing as a function of c in the ordering $\leq_H$, i.e.,

$$c_1 X \leq_H c_2 X \quad \textit{if} \quad c_1 < c_2 \ . \tag{6.14}$$

Proof. To establish (6.14), it suffices to show that $\lambda_{c_1 X}(t) \geq \lambda_{c_2 X}(t)$ for all t, but that is equivalent to

$$\frac{\lambda_X(t/c_1)}{c_1} \geq \frac{\lambda_X(t/c_2)}{c_2} \quad \text{for all} \quad t \ , \tag{6.15}$$

which we see is equivalent to (6.13) if we make the change of variables: $x = t/c_1$ and $c = c_2/c_1$. ∎

For the gamma pdf in (3.2), it is important that the sufficient conditions in Theorem 6.1 are satisfied when we make direct changes to the parameters, in an appropriate way. In particular, if we increase the mean $m_G = \nu/\mu$ by either (1) decreasing μ, while holding ν fixed or (2) increasing ν, while holding μ fixed, then G increases in the $\leq_{LR}$ ordering, which implies that G increases in the required $\leq_H$ ordering; e.g., see Problem 9.21 of [27]. Thus, increasing the mean m_1 of G in that way for the gamma agent-retention cdf causes the overall long-run average performance $\mathbf{r}(m_1)$ to increase.

7. Statistical Issues: Fitting the Functions

In this section we briefly discuss statistical issues associated with fitting our proposed model to data, but we do not analyze any data here. For a recent extensive statistical study of contact-center data, see Brown et al. [9].

Our model showing how agent retention affects contact-center performance has three elements: the performance function $r \equiv r(t)$, the agent-retention cdf $G \equiv G(t)$ and the staff-experience cdf $F \equiv F(t)$. Under our model assumptions, we have $F = G_e$ by Theorems 3.1 and 3.2, where $G_e \equiv G_e(t)$ is the stationary-excess cdf associated with G defined in (3.5). Hence, under our model assumptions, there are only two model elements to be specified r and G. However, it may be easier to estimate F directly than estimate G. Our model and the steady-state conditions for that model (assumed in Section 3) allow us to represent F by G_e. We should recognize that those assumptions might not be justified, but we proceed assuming that they are.

In this section we discuss statistical procedures to estimate the model elements from contact-center data. Our goal is to obtain estimators $\hat{r}$, $\hat{G}$ and $\hat{F}$ for the three model elements r, G and F. We should keep in mind that these quantities are all functions of the length of service t, not simple numbers.

We start by considering the performance function r. First, of course, we must specify how agent performance is to be measured. A simple measure readily available from the *automatic call distributor* (ACD) is the number of calls handled by the agent per (working) hour. With the aid of the *customer relationship management* (CRM) system and the ACD, we could instead use a measure such as revenue generated by the agent per hour. Thus, for each agent, and each time period (a day, say), we would obtain a data point (y, t), where y represents the observed performance and t representing the length of time that the agent has been employed. For any individual agent, say agent j, we can estimate agent $j's$ performance function $r_j \equiv r_j(t)$ by fitting the function r_j to the set of (y, t) pairs for that agent. Similarly, for the entire contact center, we can estimate the performance function $r \equiv r(t)$ by fitting that function to all the (y, t) pairs. Of course, this is a statistical problem. If there were a perfect fit, then we would have $y = r(t)$ for all pairs (y, t) and some function r. But we cannot nearly expect that. Instead, we statistically fit a function r to the data. That fitted function is our estimator $\hat{r}$. In the process of doing the function fitting, we should also measure the statistical validity of the relation.

Next we turn to the agent-retention cdf G. The obvious direct approach is to go into the employment records and obtain the length of service for each agent that has worked for the contact center. However, there are difficulties. First, we do not know when to start measuring. If we go back in time too far, then the data may not be representative of the current conditions of the contact center. But suppose that we select an appropriate measurement period. We

would then consider the population of all agents that started work during that interval of time. We should recognize that we would have statistical problems if we, instead, focus on the agents who worked any time during that interval (including those that started employment before the measurement interval), because there would then be a selection bias. So suppose we focus only on the agents that started work during the designated time period.

Then the natural estimator $\hat{G}$ for the agent-retention cdf G is the empirical cdf $G_n \equiv G_n(t)$, based on the sample of size n: $G_n(t)$ is the proportion of the n sampled agents that were employed for a total time less than or equal to time t. We might use a statistical smoothing technique to estimate the agent-retention pdf g from the histogram g_n associated with the empirical cdf G_n. (The histogram is essentially a probability mass function assigning mass $1/n$ to the retention time of each of the n agents. The histogram goes further by grouping the values into subintervals.)

Unfortunately, however, there are further difficulties with this direct and natural approach. In particular, we also have the problem of *censored data*: We can only accurately measure the total length of service for those agents that already have terminated employment. There may well be a significant number of agents in our sample (who started employment during our measurement interval) who are still currently employed. All we know about these agents is their current length of employment. For them, that current length of employment underestimates their ultimate, yet-to-be-determined, total length of employment. On the other hand, if we take the agents still working out of the sample, then we look only at agents completing service in the observation window, causing us to bias the estimate the other way, not counting agents with longer service times. If the number of agents currently employed is a relatively small part of the whole sample, then this difficulty can be considered minimal.

However, we anticipate that the number of agents currently employed will be a relevant part of the overall sample, so it is likely that the censored-data problem will have to be addressed. Fortunately, there are available statistical procedures to cope with censored data, via survival analysis, and in particular the Kaplan-Meier estimator [22, 32]. See Section 6 of Brown et al. [9] for applications of this analysis to analyze abandonment and waiting in contact centers.

Given that we can indeed obtain an estimator $\hat{G}$ to estimate the agent-retention cdf G, e.g., by the empirical retention cdf G_n, we can obtain an estimator $\hat{f}$ for the staff-experience pdf f by letting

$$\hat{f}(t) = \frac{1}{\hat{m}_G}\hat{G}^c(t), \quad t \geq 0 \ , \tag{7.1}$$

where $\hat{m}_G$ is an associated estimator for the mean of G (naturally taken to be the mean of $\hat{G}$)

and $\hat{G}$ is the chosen estimator for G. We obtain (7.1) by simply using the chosen estimator $\hat{G}$ and the established relation between f and G in (4.2).

We conclude this section by proposing an alternative approach that avoids two of the difficulties above: (i) going back in time and (ii) censored data because agents to be considered are still employed. To avoid these difficulties, we suggest focusing instead on the *empirical staff-experience cdf* $F_t \equiv F_t(x)$ defined in Section 1: $F_t(x)$ is the proportion of the agents working at time (day, say) t that have been employed for a length of time less than or equal to x (months, say). The obvious advantage of using the empirical cdf F_t is that it is directly observable. It is itself a directly measurable quantity. Moreover, it too can be obtained by carefully exploiting the employment records.

Our analysis in Section 3 is important because it shows how to interpret the empirical staff-experience cdf F_t. In particular, we now understand how to relate F_t to the overall long-run average staff-experience cdf F and the agent-retention cdf G. Of even greater importance, we clearly see that F_t, F, G and the estimator $\hat{G}_n$ defined above are *four different but related functions*. We would suggest estimating the desired staff-experience cdf F by a finite time-average of F_t; i.e.,

$$\hat{F}_{T,N}(x) = \frac{1}{N+1}\sum_{i=0}^{N} F_{iT/N}(x) \quad \text{for all} \quad x > 0 \ , \tag{7.2}$$

where $[0, T]$ is the selected measurement interval, with T representing the current time and 0 representing a time T units in the past, which we have decided to divide into $N + 1$ evenly spaced observation times.

We believe that it may be useful to look at the empirical cdf F_t and see how it evolves over time. We can see the evidence of changes in turnover through the evolution of F_t. Since F_t is a random variable, it is natural to smooth it by taking time averages. Hence we might look at the time average

$$\hat{F}_{u,T,N}(x) = \frac{1}{N+1}\sum_{i=0}^{N} F_{u+(iT/N)}(x) \quad \text{for all} \quad x > 0 \ , \tag{7.3}$$

as a function of u. The estimator $\hat{F}_{u,T,N}(x)$ estimates the time-average of F_t over the time interval $[u, u+T]$ as a function of u. If the (random) cdf $\hat{F}_{u,T,N}(x)$ tends to increase stochastically, in some sense, as u increases we see retention improvements over time, measured in that way.

It is natural to ask if the contact center can be regarded as being in steady-state at any given observation time. One indication of that would be that there is no systematic trend

in the statistic $\hat{F}_{u,T,N}(x)$ as the measurement starting time u changes. Assuming that the contact center can indeed be regarded as being in steady state, the estimators $\hat{F}_{T,N}(x)$ in (7.2) and $\hat{F}_{u,T,N}(x)$ are in fact direct estimators of the staff-experience cdf F. Given one of these estimators, say $\hat{F}$, we can apply statistical methods to obtain an associated estimator $\hat{f}$ for the staff-experience pdf f. We can then apply formula (4.2) to obtain an estimator $\hat{G}$ for the associated agent-retention G, namely,

$$\hat{G}(x) = 1 - \frac{\hat{f}(x)}{\hat{f}(0)} \ . \tag{7.4}$$

This procedure leads us to estimate the mean of G by $\hat{m}_G = 1/\hat{f}(0)$.

8. Transition Costs

In this section we supplement our probability model in Section 3 in order to describe the transition costs discussed in Section 1. As before, we assume that the number of agents working in the contact center is fixed at n for all time, with a new agent hired whenever a working agent departs. As before, we assume that the agent employment durations $X_{i,k}$ are IID random variables for $1 \le i \le n$ and $k \ge 1$, with $X_{i,k}$ representing the length of time that the k^{th} agent in the i^{th} agent position is employed. As before, we assume that $X_{i,k}$ is distributed as the random variable X with cdf G having finite mean $m_G = E[X]$.

Let $N_i \equiv \{N_i(t) : t \ge 0\}$ be the *renewal counting process* associated with the i^{th} position, i.e.,

$$N_i(t) \equiv \max\{k : S_{i,k} \le t\}, \quad t \ge 0 \ , \tag{8.1}$$

where

$$S_{i,k} \equiv X_{i,1} + \cdots + X_{i,k}, \quad 1 \le i \le n, \quad \text{and} \quad k \ge 1 \ , \tag{8.2}$$

with $S_{i,0} \equiv 0$ for each i.

The total number of transitions in the time interval $[0, t]$ is then

$$N(t) \equiv \sum_{i=1}^{n} N_i(t), \quad t \ge 0 \ . \tag{8.3}$$

By Proposition 3.3.1 and Theorem 3.3.4 of Ross [28], we can describe the long-run transition rate, justifying a claim made in the introduction.

Theorem 8.1. *Under the assumptions above,*

$$\frac{N(t)}{t} \to \frac{n}{E[X]} \quad \textit{with probability 1} \quad \textit{as} \quad t \to \infty \tag{8.4}$$

and

$$\frac{E[N(t)]}{t} \to \frac{n}{E[X]} \quad as \quad t \to \infty \ . \tag{8.5}$$

In other words, for each agent position, the long-run average transition rate is $1/E[X]$.

Now we consider the costs associated with each transition. In doing so, we do not do a detailed analysis. Instead, we simply assume that there is an additional set of IID random variables $Z_{i,k}$ for $1 \leq i \leq n$ and $k \geq 1$, with $Z_{i,k}$ representing the random cost associated with the k^{th} transition at the i^{th} agent position. Let the random variables $Z_{i,k}$ be distributed as a random variable Z with cdf H having finite mean $E[Z]$.

Then the total transition cost at the i^{th} agent position during the time interval $[0, t]$ is

$$C_i(t) = \sum_{k=1}^{N_i(t)} Z_{i,k} \ , \quad t \geq 0 \ , \tag{8.6}$$

and the overall *total transition cost* during the time interval $[0, t]$ is

$$C(t) = \sum_{i=1}^{n} \sum_{k=1}^{N_i(t)} Z_{i,k} \ , \quad t \geq 0 \ . \tag{8.7}$$

Under all the IID assumptions made above, for each i, the stochastic process $C_i \equiv \{C_i(t) : t \geq 0\}$ is a *renewal-reward process*, as in Section 3.6 of Ross [28]. In turn, the stochastic process, $C \equiv \{C(t) : t \geq 0\}$ is the sum of n IID renewal-reward processes. Thus, be Theorem 3.6.1 of [28], we have the following result, describing the long-run average total transition cost.

Theorem 8.2. *Under the assumptions above,*

$$\frac{C(t)}{t} \to \frac{nE[Z]}{E[X]} \quad \textit{with probability 1} \quad \textit{as} \quad t \to \infty \tag{8.8}$$

and

$$\frac{E[C(t)]}{t} \to \frac{nE[Z]}{E[X]} \quad as \quad t \to \infty \ . \tag{8.9}$$

In practice (for actual finite times t), the observed average cost $C(t)/t$ will inevitably differ from the long-run average $nE[Z]/E[X]$. Under the model assumptions, random fluctuations about the limit can be described by a central limit theorem. By Theorem 7.4.1 of [35], we obtain the following characterization. For the statement, let $N(m, \sigma^2)$ denote a random variable with a normal distribution having mean m and variance σ^2. As in (3.4), let $\Rightarrow$ denote convergence in distribution.

Theorem 8.3. *If, in addition to the assumptions above, the variances* $\sigma_X^2 \equiv Var(X)$ *and* $\sigma_Z^2 \equiv Var(Z)$ *are finite, then*

$$\frac{C(t) - \gamma t}{\sqrt{t}} \Rightarrow N(0, n\sigma^2) \quad as \quad t \to \infty \ , \tag{8.10}$$

where

$$\gamma \equiv \frac{nE[Z]}{E[X]} \quad and \quad \sigma^2 \equiv \frac{\sigma_Z^2}{E[X]} + \frac{E[Z]^2\sigma_X^2}{E[X]^3} \ . \tag{8.11}$$

As a consequence of Theorem 8.3, we see that, for large t, $C(t)$ is approximately normally distributed with mean γt and variance $n\sigma^2 t$ for γ and σ^2 in (8.11).

9. Conclusions

In this paper we have presented a mathematical framework to help think about the way management actions to increase agent job satisfaction (increasing compensation or autonomy, reducing stress, or by any other means) may increase agent retention and enhance contact-center performance. We have developed mathematical models to describe both the transition costs of turnover (Section 8) and the performance benefits of retention (Sections 2, 3 and 5). We believe that the models and analysis can be useful when combined with empirical analysis of contact-center data. The models and analysis can even help guide the empirical analysis.

Mathematical models have an automatic precision that requires careful definition. Thus the act of modelling can help us carefully define the quantities being studied. For example, the model identifies two quantities that might be confused: the length of time each agent works (modelled by the agent-retention cdf G) and the experience of the staff at any time (modelled by the staff-experience cdf F). The model also determines a precise relation between these two quantities, under assumptions. Under the model assumptions, we have shown how changes in the agent-retention cdf G will produce corresponding changes in the staff-experience cdf F. It is natural to next investigate if these relationships are seen in practice.

Given that management actions may significantly affect agent job satisfaction, with some actions acting positively, but possibly others acting negatively (e.g., pervasive monitoring), it is desirable to investigate how these actions actually do affect retention, staff experience and performance. By measuring all these quantities over time, management can learn about the costs and benefits of those management actions.

The modelling and analysis raise important empirical issues. For example, we see that it is natural to ask how performance might best be quantified. Moreover, for an appropriate

quantification, our model leads us to ask if performance can indeed be regarded as an increasing function of experience and, if so, what is the shape of the function? And how much performance benefits are gained by increasing staff experience? More fundamentally, we suggest considering that actions to increase agent job satisfaction might be cost-effective. It is possible that such measures can be *win-win-win* actions; all the parties - the agents, the company and the customers - might simultaneously benefit. If such win-win-win opportunities exist, it would be desirable to find them.

10. Acknowledgments

I thank Michael E. Sisselman for stimulating discussions that motivated this work. I thank Peter Bamberger and Michal Borin of the Technion for insightful comments and for directing me to research literature related to job satisfaction, retention and performance in contact-centers.

References

[1] Abate, J. and Whitt, W. Transient behavior of regulated Brownian motion, I; starting at the origin. *Advances in Applied Probability* 19 (1982) 560–598.

[2] Abelson, M. A. and Baysinger, B. D., Optimal and dysfunctional turnover: toward an organizational level model. *Academy of Management Review* 9 (1984) 331–341.

[3] Asmussen, S. *Applied Probability and Queues*, second edition, Springer, 2003.

[4] Barlow, R. E. and Proschan, F., *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, 1975.

[5] Bartholomew, D. J., Forbes, A. F. and McClean, S. I. *Statistical Techniques for Manpower Planning*, second ed., Wiley, 1991.

[6] Batt, R. Strategic segmentation in front line service: matching customers, employees and human resource systems. *International Journal of Human Resource Management* 11 (2000) 540–561.

[7] Batt, R. Managing customer services: human resource practices, quit rates and sales growth. *Academy of Management Journal* 45 (2002) 587–599.

[8] Bliss, W. G. Cost of employee turnover. *The Advisor.*
Available at: http://www.isquare.com/turnover.cfm

[9] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Zeltyn, S., Zhao, L. and Haipeng, S. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association* (JASA), to appear.

[10] Bryant, D. T. and Niehaus, R. J. (editors) *Manpower Planning and Organizational Design*, Plenum Press, 1978.

[11] Cleveland, B. and Hash, S. (editors) *Call Center Agent Motivation and Compensation*, Call Center Press, ICMI, Annapolis, MD, 2004.

[12] Coffman, E. G., Jr., Flatto, L. and Whitt, W. Stochastic limit laws for schedule makespans. *Stochastic Models* 12 (1996) 215–243.

[13] Cordes, C. L. and Dougherty, T. W., A review and integration of research on job burnout. *Academy of Management Review* 18 (1993) 621–656.

[14] Cotton, J. L. and Tuttle, J. M., Employee turnover: a meta-analysis and review with implications for research. *Academy of Management Review* 11 (1986) 55–70.

[15] Feldmann, A. and Whitt, W. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation* 31 (1998) 245–279.

[16] Gans, N., Koole, G. and Mandelbaum, A. Telephone call centers: tutorial, review and research prospects. *Manufacturing and Service Operations Management* (M&SOM), 5 (2003) 79–141.

[17] Gans, N. and Zhou, Y.-P. Managing learning and turnover in employee staffing. *Operations Research* 50 (2002) 991–1006.

[18] Glebbeek, A. C. and Bax, E. H., Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal* 47 (2004) 277-286.

[19] Grinold, R. C. and Marshall, K. T. *Manpower Planning Models*, Elsevier North-Holland., 1977.

[20] Holman, D. Employee wellbeing in call centres. *Human Resource Management Journal* 12 (2002) 35–50.

[21] Holman, D. Call centres. Chapter 7 in *The New Workplace: A Guide to the Human Impact of Modern Work Practices*, D. J. Holman, T. D. Wall, C. W. Clegg, P. Sparrow and A. Howard (editors), Wiley, 2003.

[22] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*, second edition, Wiley, 2002.

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

[24] Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J. and Erez, M., Why people stay: using job embeddedness to predict voluntary turnover. *Academy of Management review* 44 (2001) 1102–1121.

[25] Müller, A. and Stoyan, D. *Comparison Methods for Stochastic Models and Risks*, Wiley, 2002.

[26] Quiñones, M. A., Ford, J. K. and Teachout, M. S., The relationship between work experience and job performance: a conceptual and meta-analytic review. *Personnel Psychology* 48 (1995) 887–910.

[27] Ross, S. M. *Stochastic Processes*, second edition, Wiley, 1996.

[28] Ross, S. M. *Introduction to Probability Models*, eighth edition, Academic Press, 2003.

[29] Ruyter, K., Wetzels, M. and Feinberg, R., Role stress in call centers: its effects on employee performance and satisfaction. *Journal of Interactive Marketing* 15 (2001) 23–35.

[30] Singh, J., Goolsby, J. R. and Rhoads, G. K., Behavioral and psychological consequences of boundary spanning burnout of customer service representatives. *Journal of Marketing Research* 31 (1994) 558–569.

[31] Sisselman, M. E. and Whitt, W. Preference-based routing. SeatLink, Inc., and Columbia University, 2004.
Available at: http://www.columbia.edu/∼ww2040/recent.html

[32] Tableman, M. and Kim, J. S. *Survival Analysis using* ***S***, Chapman and Hall, 2004.

[33] Whitt, W. Approximating a point process by a renewal process, I: two basic methods. *Operations Research* 30 (1982) 125–147.

[34] Whitt, W. The renewal-process stationary-excess operator. *Journal of Applied Probability* 22 (1985) 156–167.

[35] Whitt, W. *Stochastic-Process Limits*, Springer, 2002.

[36] Witt, L. A., Andrews, M. C. and Carlson, D. S., When conscientiousness isn't enough: emotional exhaustion and performance among call center customer service representatives. *Journal of Management* 30 (2004) 149–160.

[37] Wolff, R. W. *Stochastic Modelling and the Theory of Queues*, Prentice-Hall, 1989.

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