

Redefining the service level in call centers

Ger Koole

Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands

17th June 2005

Abstract

We propose a new waiting time metric for call centers that circumvents some of the problems that the standard way of defining service level has.

As any other department in a company the performance of a call center is measured by looking at certain *performance indicators* (PIs). In the case of a call or contact center they are usually related to agent effectiveness and efficiency and to customer satisfaction. The most common PIs are agent productivity (what percentage of time is the agent handling calls, i.e., not available to take new calls), the fraction of abandonments (which percentage of callers abandons while waiting, thus before speaking to an agent), and what is often called the *service level* (SL): the percentage of callers that has to wait shorter than a specified amount of time (called the acceptable waiting time, or AWT). Taking this waiting time metric means making several choices: how many seconds waiting is acceptable, how should I count callers that abandon, which percentage is acceptable. A discussion on how to make these choices can be found in Cleveland & Mayben [1, p. 33–40].

In this paper we pose ourselves the question: is the percentage of callers that has to wait shorter than a specified amount of time (the service level) the right PI to measure customer satisfaction, i.e., is this the right waiting time metric? The reason to pose this question is the fact that it has two major disadvantages:

- The SL gives no information on how long callers that have exceeded the AWT still have to wait;
- The SL stimulates managers to give priority to callers who have not yet reached the AWT, thereby increasing even more the waiting time of callers that have waited longer than the AWT.

Thus we have no information on callers that passed the AWT and we are stimulated not to care about them. Let us illustrate both disadvantages in a numerical way.

We consider a call center with a single type of calls, Poisson arrivals and service times (talk time plus wrap up) that are approximately exponentially distributed. Callers are served in the order of arrival. This situation can be modeled by the Erlang C queueing system. The arrival rate is λ , the average service time is $\beta = \mu^{-1}$ (both counted in minutes), and there are s servers or agents. The AWT is usually 20 seconds, thus 1/3 minute. We

first show that for a given excess probability the average waiting time in excess of the AWT can take any value. Take $\lambda = 1$, $\beta = 8$, and $s = 11$. Then the SL is a little more than 78% (using the well known solution for the Erlang formula). Now take $\lambda = 2$ and $\beta = 4$: then the SL is 81%. That both cases give almost the same SL can be explained as follows. In both cases the load $a = \lambda\beta$ is equal, and therefore the blocking probability (commonly written as $C(s, a)$) is equal for both cases. Most of the SL is reached by customers that do not wait at all ($C(s, a) = 0.24$), because $\text{AWT} \ll \beta$. Therefore the SL does mostly depends on product a , less on the separate values λ and β . This does not hold for the residual waiting time for customers that exceeded the AWT. Denote the waiting time in the Erlang C queue by W_q . It is well known that $W_q|W_q > 0$ is exponentially distributed with parameter $s\mu - \lambda$. Thus the residual waiting time $W_q - \text{AWT}|W_q > \text{AWT}$ is also exponentially distributed, with expectation $1/(s\mu - \lambda) = \beta/(s - a)$. Thus from one case to the next the residual waiting time doubles. By taking $\lambda = 1/k$ and $\beta = 8k$ for increasing k the SL converges to 76% while the residual waiting time given $W_q > \text{AWT}$ increases linearly in k . This is an undesirable situation, the waiting times of customers that wait longer than the AWT should somehow be expressed in the way waiting times are measured.

The arguments just given show that the SL as it is defined usually gives a wrong or at least incomplete image of the customer satisfaction. A second, related disadvantage of using the SL becomes clear when we schedule calls rationally, i.e., in such a way that the SL is optimized. For this we assume that we can choose which call to serve next. In this case it is no longer optimal to serve calls in order of arrival. It is intuitive clear and easily shown using a coupling argument that is optimal to serve the ‘oldest’ call that has not yet reached the AWT, and that is never optimal to serve the calls that have exceeded the AWT! To illustrate the “SL improvement” for this policy with respect to the usual first-come-first-served (FCFS) order a simulation study would have to be executed: analytical solution methods are not available. However, a simpler rule in which calls are ‘dropped’ (i.e., they never receive service) in a random way can easily be analyzed using the Erlang formula. It also shows a considerable improvement, compared to FCFS. Take $\lambda = 10$, $\beta = 2$, $\text{AWT} = 1/3$, and $s = 23$. Then $\text{SL} = 75\%$. Now drop every customer with a probability of $1/20$. The calls that are served arrive according to a Poisson process with parameter $\lambda = 9.5$. For the customers that are served $\text{SL} = 85\%$. Because 5% is blocked, we find $0.95 \times 0.85 = 0.81$, and thus a SL of more than 80%.

Our conclusion is that the standard waiting time metric is badly chosen. Next we propose a new way to summarize the information on waiting times that is equally simple as the SL and where rational behavior does not lead to unwanted behavior. Before doing so, we should realize that mathematically speaking the “best” waiting time metric is the waiting time distribution W_q itself. In the Erlang C model, it consists of two parameters: the delay probability, and the parameter of the exponential distribution of $W_q|W_q > 0$. When using other models to represent waiting in call centers other representations of W_q are necessary. Here however we focus on representing the information on waiting times by a single number. This is simple, and it facilitates comparisons.

First we formalize the concept of waiting time metric. Assume that the waiting time metric \mathcal{W} is the expected value of a function f of the waiting time W_q of an arbitrary

customer: $\mathcal{W} = \mathbb{E}f(W_q)$. We interpret $f(t)$ as the cost of waiting t time units, and thus \mathcal{W} should be as small as possible. SL fits in this framework as follows. SL should be as high as possible, and thus $100 - \text{SL}$ should be as small as possible. For $\mathcal{W} = 100 - \text{SL}$ the function f is given by $f(t) = 100\mathbb{I}\{t > \text{AWT}\}$.

From the previous discussion it is clear that we should have a waiting time metric for which it is optimal to schedule the longest waiting call first. From a simple coupling argument it follows that this is the case if f is convex; it is the only optimal action if f is strictly convex.

A waiting time metric that is convex and that is also often used as PI is the expected waiting time, in call centers known as the *average speed of answer* (ASA). However, the ASA is rarely used as the major waiting time metric, because it does not take into account the variability of the waiting time. E.g., two calls waiting 20 seconds is often considered preferable to one call waiting 40 and another call 0 seconds, certainly if $\text{AWT} = 20$. The SL does differentiate between these two cases, but it does not penalize waiting long, because f corresponding to it does not increase after AWT.

We introduce a new waiting time metric $\mathcal{W} = \mathbb{E}f(W_q)$. It should have the following properties:

- \mathcal{W} is simple to interpret, any manager can understand its meaning;
- \mathcal{W} takes the AWT into account;
- $f(t)$ is increasing for $t > \text{AWT}$;
- \mathcal{W} consists of a single number (i.e., f is one-dimensional);
- f is convex.

A waiting time metric that satisfies all these conditions is the *average excess* (AE), defined by $f(t) = (t - \text{AWT})^+$. It gives the average time waiting exceeds the AWT. It is also easily used for waiting time calculations, given the simple Erlang C model discussed earlier. From the fact that $W_q|W_q > 0$ is exponentially distributed with parameter $s\mu - \lambda$, it follows that $W_q|W_q > \text{AWT}$ has the same exponential distribution. Therefore

$$\mathbb{E}(W_q - \text{AWT})^+ = \frac{C(s, a)e^{-(s\mu - \lambda)\text{AWT}}}{s\mu - \lambda}.$$

Its use is illustrated in Table 1. The first three line concern the situations where the SL does not change much, but ASA and AE do, for reasons explained earlier. Note that when there are calls that never receive service then $\text{AE} = \infty$. In reality customers abandon if they receive such a bad service; this is one of the reasons why the Erlang C model is not always a good choice for call center modeling (see Gans et al. [2] for an overview of call center modeling).

The next two lines of Table 1 are situations where ASA is roughly constant, but where SL and AE do vary significantly. This difference can be explained by the variability of the waiting time distribution W_q . The last two lines show the effect of changing the AWT.

λ	β	s	AWT	SL	ASA	AE
2	4	11	20	81%	19	15
1	8	11	20	78%	39	35
0.5	16	11	20	77%	78	74
10	2	24	20	75%	9	5
40	0.5	22	20	85%	9	2
10	1	12	20	77%	13	7
10	1	12	40	88%	13	4

Table 1: Performance indicators (ASA and AE in seconds) for several parameter value combinations (λ and μ in minutes, AWT in seconds) for the Erlang C model

References

- [1] B. Cleveland and J. Mayben. *Call Center Management on Fast Forward*. Call Center Press, 1997.
- [2] N. Gans, G.M. Koole, and A. Mandelbaum. Operational models of telephone call centers: a tutorial and literature review. *Manufacturing & Service Operations Management*, 2003. To appear.