Advanced Topics in Call Center Staffing

Call center calculators such as EasyErlang are important productivity tools in planning and managing the key resources in your call center, namely the number of customer service representatives (CSRs) and telephone lines. These tools estimate resource requirements, recommend staffing and scheduling, and forecast key operational metrics. However, some call center staffing calculators take a somewhat simplified view of the operational structure of the call center, and some analysis and preprocessing is required in order to accommodate some real-world situations that you are likely to encounter. In this article we discuss three such cases and how to address them:

A common practice in call centers is to route incoming calls depending on the type of customer or service request. The idea is to Instead of all callers recewaiting Each call is assigned to one of several quueu and route them to specialized groups or individuals, where they queue separately. However, despite what may be intuitively logical, this practice is usually less efficient. With a smaller number of calls and servers per queue, the statistical distribution of service traffic decreases the efficiencies in each queue, thereby decreasing the overall efficiency. Table 4 compares the anticipated staffing and utilization rates for a one, two and three queue help desk.

Table 4. Impact of Multiple Queues

Calls/hour	100	100	100
Number of Queues	1	2	3
Calls/ Queue /hour	100	50	34
Staff/Queue	19	11	8
Total staff	19	22	24
Utilization	79%	68%	64%

Staffing for Multiple Queues

The majority of call center planning tools are based on the Erlang model. This model assumes that all calls arrive into a single queue and are served by a single pool of telephone agents. In reality, many call centers are comprised of multiple queues assigned to different products or services. Each queue, therefore, may be experiencing a different call load and performance characteristics, and will be required to meet the expectations of a different customer group.

The method to calculate staffing requirements for a multi-queue call center depends on the configurations of the queues.

Disjoint Queues

Disjoint queues are separate call center queues. Each queue utilizes dedicated telephone resources and agents answer only calls arriving in their queue. Disjoint queues represent the simplest form of multiple queues in that each queue is independent and has no effect on the other queues; in essence, each queue behaves as a separate call center.

In a call center with disjoint queues, you simply calculate the resources for each of the queues separately. In most cases, you can estimate the resources by adding the requirements of the individual queues.

Joint Queues

The individual queues in many call centers are not purely disjointed. Instead, they operate separately until a queue exceeds its capacity, and calls overflow to a less busy “backup” queue. From that point on, the two queues are joint and act like a single larger queue.

When calculating the resources and the performance characteristics of joint queues, the call volume is the sum of the call volumes of the individual queues. However, the average handling time (AHT) of calls in different queues is

usually different and therefore must be adjusted before entered into the staffing calculator. It is incorrect simply to average the AHT of the joint queues. For example, if the AHT of queue Q1 is 100 seconds and the AHT of queue Q2 is 200 seconds, the AHT of the combined queues is not 150 seconds. Not only is it mathematically incorrect to average numbers that are averages themselves, but also the number of calls of each type (i.e. precisely routed to individual queues) is also different. Therefore, the new AHT should be a weighted average that is proportionally adjusted to the call volume of each AHT. Consider the following call statistics:

Queue	Q1	Q2
Calls	5,000	3,000
AHT	100	200

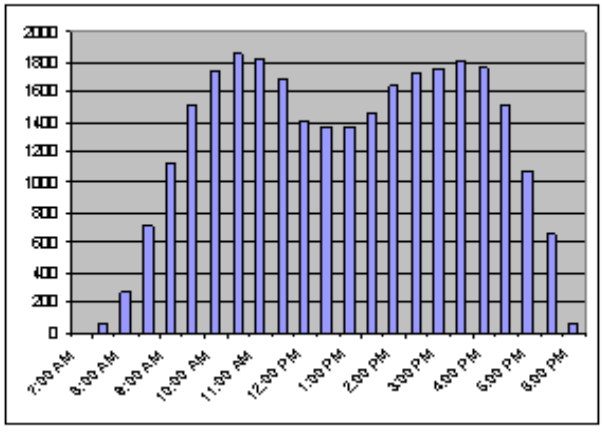
The AHT using weighted average calculation is:

$$\frac{(5000 \times 1000) + (3000 \times 200)}{5000 + 3000} = 137.5$$

The output of the call center calculator: staffing, service levels, agent utilization, abandonment rate and so forth, will be the same for both queues at peak times, when overflow occurs. You can still use the calls per hour and AHT for each of the queues to ascertain service levels during normal operation when the queue is not fully occupied.

If agents in the backup queue are less trained or equipped to handle overflow calls, their AHT is likely to be longer than the AHT of the agents in the original queue. In this case, a better approach is to determine the maximum call handling capacity of each queue (EasyErlang calculates the maximum capacity and the additional capacity required to meet peak load), and the estimated number of overflow calls. Use the same weighted average method to calculate the resources and performance of the backup queue during peak time.

Staffing for Variable Load



The call load in a typical call center varies from hour to hour and from day to day, sometimes from season to season. The figure on the right shows changes in daily call load in a typical call center. The commonsense approach that is usually employed by call center planners is to calculate staffing separately for the highest load and the low load periods. A call center calculator can estimate the service level for each of these periods, but what is the overall daily service level of the entire call center?

To calculate the combined service level for several load levels or for different shifts, one cannot simply average the individual service levels. For example, if the service levels for three shifts are 89%, 91% and 84% of calls answered in 20 seconds, the overall service level is not 88%. Rather, each service level has to be computed considering the total number of calls offered and the number of callers that have experienced each service level, similar to the method we used to calculate AHT for different queues.

Period	P1	P2	P3
Calls Offered	200	500	300
Service Level	89%	91%	84%

$$\frac{200}{1000} \times 89\% + \frac{500}{1000} \times 91\% + \frac{300}{1000} \times 84\% = 86.3\%$$

The same method is used to convert daily service levels to weekly or monthly service levels.

This method also provides a more pragmatic way to ensure that the call center meets service level targets without overstaffing. Consider a call center that averages 500 calls per hour during peak time, but only 280 calls per hour during the rest of the day. Staffing for the peak period, assuming a target service level of 80% in 20 seconds, will require 30 agents. Staffing for the low volume period will require only 18 agents. As the table below shows (see Row 1), when staffed properly, the service levels in each of the two periods, as well as the overall daily service level, exceed the required performance target.

	Period I		Period II		Overall
	CSRs	**SL**	**CSRs**	**SL**	**SL**
(1)	30	85.66%	18	85.14%	85.48%
(2)	29	78.13%	18	85.14%	80.65%
(3)	30	85.66%	17	75.05%	85.13%
(4)	31	90.77%	17	75.05%	85.13%
(5)	29	78.13%	19	91.42%	82.90%

If we reduce the headcount in Period I by one CSR, the service level in that period will drop to 78% (Row 2), which is not significantly worse than the target of 80% and is probably tolerable, especially if peak time is short in duration. Note the overall service level meets the target of 80%.

Row 3 shows what happens if we decrease the headcount of the low volume period. The impact on the service performance during that period is more significant, yet the overall service level is still above the target.

Interestingly enough, moving the headcount that was “saved” from one period to the other improves the service level during that period, as expected, but the overall service level is actually lower than the one achieved using the optimal allocation (see Rows 4 and 5).

We conclude that staffing for very dynamic load conditions while maintaining service level agreement and minimizing inefficiencies can be challenging. As the examples show, this type of planning does benefit from overly

precise measurement of call load in frequent intervals. Instead of adjusting staffing for each period in the day, call center planners and managers should focus on marinating overall good performance.

Shrinkage

The term *shrinkage* is used to identify the portion of work time that a CSR is unavailable to handle calls. This includes *planned shrinkage* such as breaks, meetings, training, research and special projects, vacation, and unscheduled shrinkage such as sick time.

Shrinkage can easily accounts for more than 20% of agents' paid time. Ignoring shrinkage will cause a significant understaffing, especially during peak load time. How can a call center planner adjust staffing to compensate for this productivity loss? Call center planners are often confused by this.

The simplest answer appears to be to increase the headcount by the amount lost to shrinkage. If the staffing calculator determines that current service level and call load require 60 agents, and the anticipated shrinkage is 25%, then: 60 x 1.25 = 75.

Or, you could say that the staff of 60 is available only at 75% capacity because of the 25% shrinkage. Therefore, in order to achieve 100% capacity, the headcount should be 60 / 0.75 = 80.

The difference between the two methods is 5 CSRs. Moreover, the agents we need to add in order to account for the shrinkage would also be subject to the same shrinkage, accounting, in this case, to nearly 2 additional CSRs: 5 / 0.75 = 6.7, or 7 CSRs.

[what's the correct method?] [text re: how does EE calculate?]

The impact of shrinkage in small call centers is not very significant. However, in large call centers, where the shrinkage compensation is significant, an more accurate adjustment would be to account for the additional shrinkage as well.

Further Reading

- The Origin of the 80/20 Rule
- Are Abandoned Calls Important?
- Service Level Calculations
- Advanced Topics in Call Center Staffing
- Introduction to Traffic Modeling and Resource Allocation in Call Centers
- Benchmarking in Call Centers
- Does Self-Help Really Help?
- Service Level Elasticity
- An Alternative to the Erlang Traffic Model