

## Module 3 – Establishing and Using Service Level and Response Time Objectives

### 3.1 Definitions and Use of Service Level and Response Time

#### Key Points

- **Service level and response time objectives are concrete targets used for resource planning and budgeting.**
- **For planning and management purposes, all contacts must be categorized into either service level or response time buckets:**
  1. **Contacts that must be handled when they arrive are measured by service level. Service level is “X percent of contacts answered in Y seconds.”**
  2. **Contacts that can be handled at a later time are measured by response time. Response time is “100 percent of contacts handled within N days/hours/minutes.”**

#### Explanation

The principle of service level (sometimes generally referred to as accessibility or service objective) is at the heart of effective contact center management. Without service level objectives, answers to many important questions would be left to chance, e.g., How accessible is the contact center? How many staff are needed? How does the center compare to the competition? Is it prepared to handle the response to marketing campaigns? How busy are the agents going to be? What are the costs going to be?

Service level and response time objectives tie the resources needed to the desired results. They measure how well transactions are getting “in the door” and to agents so that they can get on with the business at hand, and they are stable, concrete targets for resource planning and budgeting.

#### **Definition of Service Level**

Service level is defined specifically as: “X percent of contacts answered in Y seconds,” e.g., 90 percent answered in 20 seconds or 90/20.

Service level is NOT:

- Average speed of answer, although average speed of answer is a related measure that is derived from the same set of data.
- X percent of all contacts answered, which is the inverse of abandonment. For example, it is an incorrect assumption that a 97 percent answer rate would inherently mean a 3 percent abandonment rate.
- Longest delayed call.
- X percent of contacts answered in Y seconds and the remainder of contacts answered in Z seconds. Taking this approach is essentially a service level objective of 100 percent answered in Z seconds, which will result in different staffing requirements from those required to achieve X percent answered in Y seconds. You can't have two service level objectives for the same queue.

### **Definition of Response Time**

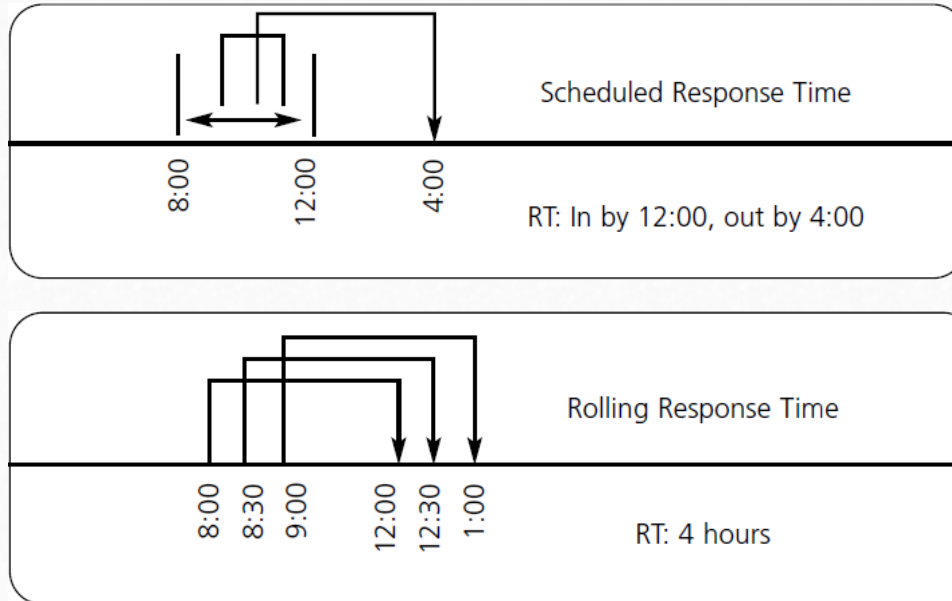
Response time is the equivalent of service level for transactions that do not have to be handled the moment they arrive. Response time is defined as "100 percent of contacts handled within N days/hours/minutes," e.g., all email will be handled within four hours or all faxes will be responded to within 24 hours.

Response time should not be confused with other uses of the word "response." Response can also be used to indicate:

- **Automated reply:** This is a system-generated response that automatically sends a reply to the customer acknowledging that a message they sent was received and informing them of when to expect a response. This establishes appropriate expectations and minimizes phone calls or other additional contacts inquiring about the status of the original message.
- **Agent-assisted response:** This refers to the response the customer receives when the transaction is actually handled by an agent. The time that elapses between the customer's original message and the contact center's response is measured as response time.
- **Resolution:** This is a measure of when the problem or issue is actually resolved and is used in environments where the contact center's initial response may not fully resolve the issue. For example, in a technical support environment, additional research may be necessary; the problem is "resolved" when the matter is handled to completion and the case is closed.

## Two Types of Response Time

There are two types of response time, scheduled and rolling.



Scheduled response time, like a dry-cleaning service, is geared around blocks of time. For example, you may commit to handle all messages received up to noon by 5 p.m., and to respond to messages received between noon and 5 p.m. by 10 a.m. the next morning.

Rolling response time is hinged on the specific times each message arrives. For example, if you establish a four-hour response time, a customer who sends a message at 9:03 a.m. should get a response by 1:03 p.m., and one who sends a message at 9:12 a.m. should receive a response by 1:12 p.m.

### Categorization of Contacts

Contacts that must be handled when they arrive require a service level objective, and those that can be handled at a later time require a response time objective:

| CATEGORIZING CUSTOMER INTERACTIONS                                                                                                                                                                                                                         |                   |                   |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|-------------------|
|                                                                                                                                                                                                                                                            | Use Service Level | Use Response Time |
| Inbound phone calls                                                                                                                                                                                                                                        | X                 |                   |
| Outbound phone calls*                                                                                                                                                                                                                                      |                   | X                 |
| Email*                                                                                                                                                                                                                                                     |                   | X                 |
| Social – real-time**                                                                                                                                                                                                                                       | X                 |                   |
| Social – deferred*                                                                                                                                                                                                                                         |                   | X                 |
| SMS                                                                                                                                                                                                                                                        | X                 |                   |
| Web chat                                                                                                                                                                                                                                                   | X                 |                   |
| Web call-me-now                                                                                                                                                                                                                                            | X                 |                   |
| Web call-me-later                                                                                                                                                                                                                                          |                   | X                 |
| Web click-to-talk                                                                                                                                                                                                                                          | X                 |                   |
| Fax                                                                                                                                                                                                                                                        |                   | X                 |
| Postal mail                                                                                                                                                                                                                                                |                   | X                 |
| Video calls                                                                                                                                                                                                                                                | X                 |                   |
| Walk-in customers                                                                                                                                                                                                                                          | X                 |                   |
| *These contacts use response time if they can be deferred. Some may require handling when they arrive.<br>**Social interactions are varied, and are broadly categorized here as those that should be handled as they occur and those that can be deferred. |                   |                   |

### The Differences Between Service Level and Response Time

In application and function, service level and response time are similar. Both provide the necessary linkage between base staff required and the desired results. However, there are important differences between service level and response time. Most notably, base staff calculations are different. Service level is used in situations with randomly arriving or peaked traffic, and requires Erlang C or computer simulation. Response time contacts can be held for later processing, and thus can rely on more traditional methods of industrial planning.

### When Response Time Becomes Service Level

There is a point at which response time objectives are so quick that they belong in the service level category. For example, to raise the bar of customer service, an organization may decide to improve response time objectives for customer email from four hours to within 15 minutes. Most queuing experts agree that with an objective of less than an hour, base staff requirements should be calculated

using either Erlang C or computer simulation programs that account for the random arrival of contacts.

Some assumptions and definitions must also change. Service level, when applied to phone calls, is based on when the calls reach agents and the conversations begin. But in the case of email, customer perceptions are geared around when they receive the responses. If you want to respond to email messages within 15 minutes and they require an average five minutes of handling time, they must reach agents 10 minutes after customers send them. That becomes the threshold for which staffing requirements must be calculated.

### Example Objectives

Example service level objectives:

| EXAMPLE SERVICE LEVEL OBJECTIVES             |                                                  |
|----------------------------------------------|--------------------------------------------------|
| General Comparisons                          | Service Levels<br>(X percent answer / Y seconds) |
| Emergency services                           | 100/0                                            |
| Service level objectives that are "high"     | 90/20, 85/15, 90/15                              |
| Service level objectives that are "moderate" | 80/20, 80/30, 90/60                              |
| Service level objectives that are "modest"   | 70/60, 80/120, 80/300                            |

Example response time objectives:

| EXAMPLE RESPONSE TIME OBJECTIVES*                                                                                                                                                                                  |                           |                                           |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|-------------------------------------------|
|                                                                                                                                                                                                                    | Common                    | High End of Range                         |
| Customer email                                                                                                                                                                                                     | Within 24 hours           | Within 1 to 4 hours                       |
| Customer voicemail                                                                                                                                                                                                 | Within 24 hours           | Within 1 to 4 hours                       |
| Social – deferred**                                                                                                                                                                                                | Within 24 hours           | Within 1 to 4 hours                       |
| Postal mail                                                                                                                                                                                                        | Processed within one week | Processed within one to two business days |
| * These should be considered examples only; according to industry surveys, there is a wide disparity in response time objectives.<br>** This refers to social interactions that do not require immediate response. |                           |                                           |

When establishing and assessing either service level or response time objectives, remember that it is not just how high the overall stated objectives are, but how consistently they are hit throughout the day.

### Why Service Level?

Why service level and not percent answered, percent abandoned, average speed of answer or other alternatives? Because "X percent answered in Y seconds" gives the clearest indication of what customers experience when they attempt to reach the organization. And service level is the most stable measurement of the queue.

Average speed of answer (ASA) is a close cousin of service level and is derived from the same set of data. But a big downside to using ASA is that it is often misinterpreted. Most of us tend to assume that the average lies somewhere in the middle of a set of data, or that average represents a "typical experience." Not so with ASA. It is mathematically correct, but does not represent what happens to individual callers.

ASA has its uses, so don't throw it out. For example, ASA is an important variable when calculating the load that your systems must carry. Further, if service level is so bad that zero percent of calls are answered within Y seconds, ASA is a practical alternative to service level. But service level is usually a more reliable and more telling measure of what callers experience. You know exactly what happens to the percentage of customers you define. If you do use ASA reports, just remember, ASA is not a bell curve — longest waits are far beyond what those numbers suggest, and it's important to know what is really happening.

What about customers who give up on a phone call, chat, or similar contact, while waiting for an agent? Looking solely at abandonment rates as a measure of whether staffing levels were appropriate can be highly misleading. The point is not to ignore abandonment. A high abandonment rate is probably a symptom of significant staffing problems. But a low abandonment rate doesn't necessarily mean everything is fine. Further, if abandonment is beyond acceptable, what are you going to do? You are going to look at instances when it's out of whack, and why. You will likely run smack into a low service level. When service level is appropriate, abandonment tends to take care of itself.

Service level also has a direct impact on channel switching, which occurs when customers begin trying different access alternatives. For example, if they encounter an extended wait, they might use self-service alternatives, try different telephone numbers or IVR menu selections (which is, technically, the same channel but a different routing alternative), switch from chat to phone or vice-versa, send a targeted message to the company through a social channel, or throw up their hands and send a message to the world at large via social media (Twitter message: *Have been waiting on hold for 20 minutes now to reach @XYZco. #Fail!*). Or, they might try simultaneous (parallel) contacts and see which generates a response first.

All of this brings up a very important point: Different callers have different experiences with your contact center, even if they are part of the same set of data measured by service level, ASA and other reports. Why? Random workload arrival. Because of this reality, you will need an understanding of what happens to different callers. At a high level, service level is the single best measure of these queue experiences. In later chapters, we'll also look at other measures that will fill the gaps.

Excerpt from *Call Center Management on Fast Forward* Third Edition by Brad Cleveland, ICMI, 2012.

## 3.2 Alternative Service Level Calculations

### Key Points

- There are a number of alternative methods contact center technology may use to calculate service level. Four of the most common include:
  1.  $(\text{Calls answered in Y seconds} + \text{calls abandoned in Y seconds}) \div (\text{total calls answered} + \text{total calls abandoned})$
  2.  $\text{Calls answered in Y seconds} \div \text{total calls answered}$
  3.  $\text{Calls answered in Y seconds} \div (\text{total calls answered} + \text{total calls abandoned})$
  4.  $\text{Calls answered in Y seconds} \div (\text{total calls answered} + \text{calls abandoned after Y seconds})$

### Explanation

There are a number of alternative methods contact center technology may use to calculate service level. With some systems, the preferred calculation can be specified. In any case, it is important to know which calculation is being used since each one handles abandoned calls differently. Not all service level contacts can be abandoned by the customer (such as when using service level for emails that must be handled as they arrive); therefore, the below calculations primarily apply to incoming calls.

Here are the most common formulas for calculating service level:

1.  **$(\text{Calls answered in Y seconds} + \text{calls abandoned in Y seconds}) \div (\text{total calls answered} + \text{total calls abandoned})$** : For most situations, this alternative is preferred because the calculation includes all of the traffic received by the automatic call distributor (ACD). Calls that abandon before the objective positively affect service level. This calculation provides a complete picture of what is happening since it takes all calls into consideration.

- 2. Calls answered in Y seconds ÷ total calls answered:** This alternative only considers answered calls, and therefore is not a good reflection of all activity. Abandonment is entirely ignored. This calculation is not generally recommended.
- 3. Calls answered in Y seconds ÷ (total calls answered + total calls abandoned):** This alternative tends to be the least popular among contact center managers because all calls that abandon negatively affect service level, even those that abandon before the objective. Canada-based consultant Cheryl Odee Helm, who has made contact center reporting a focus of her practice, recommends that this measure may be appropriate in situations where calls enter a queue after they hear a delay announcement. She does not recommend this calculation in settings where callers enter a queue before they hear the announcement.
- 4. Calls answered in Y seconds ÷ (total calls answered + calls abandoned after Y seconds):** With this calculation, abandoned calls only impact service level negatively if they happen after the Y seconds specified. Calls that abandon before the objective do not affect service level. Consequently, this is a way to avoid getting “penalized” by callers who abandon quickly without ignoring abandoned calls altogether. This is an acceptable approach.

### **Blocked Calls**

When calls are blocked (receive busy signals), they are generally not included in service level reports. Since those calls would have increased workload had they been handled, busies have the effect of making service level reports look better than they should.

### 3.3 Choosing Service Level and Response Time Objectives

#### Key Points

- **There is no such thing as an “industry standard” service level because the optimum service level is affected by:**
  - **The value of an interaction**
  - **Labor costs**
  - **Telecommunications costs**
  - **The seven factors of customer tolerance**
  - **The organization’s unique customer access strategy**
  - **The organization’s desire to differentiate products or services by the level of service provided**
- **The appropriate service level or response time objective for a given situation will, most importantly, meet organizational goals and customer expectations.**
- **Alternatives for choosing service level and response time objectives include:**
  - **Listen to the voice of the customer**
  - **Go with a “middle-of-the-road” objective, e.g., 80/20**
  - **Relate to the competition – match or go higher/lower**
  - **Adjust to minimize abandonment**
  - **Use a combined approach**

#### Explanation

There is generally no "industry standard" service level that you can hang your hat on. (There are some exceptions, e.g., service levels for many utilities are regulated.) The optimum service level is affected by a host of factors, including the value of the contact, fully loaded labor costs, telecommunications costs, customer tolerances, the organization's unique customer access strategy and the desire and commitment to differentiate through service. An industry standard would have to be based on organizations having the same values for these things.

The appropriate service level or response time objective for a given situation will, most importantly, meet organizational goals and customer expectations. The best objective is one that:

- Meets customers' needs and expectations
- Keeps abandonment at acceptable levels
- Minimizes agent burnout and errors
- Minimizes expenses
- Maximizes revenue
- Supports the organization's mission and brand
- Is understood and supported by senior management

### **Approaches for Choosing Objectives**

- **Listen to the voice of the customer:** Customer feedback gathered through conversations with agents, through customer surveys, in focus groups, and in social media comments and ratings provides valuable input when choosing your objectives. You will also want to think through the seven factors of customer tolerance.

While it's always a good idea to know what your customers expect, random call arrival means that different customers have different experiences with your contact center. Even for a relatively modest service level such as 80 percent answered in 60 seconds, more than half of the callers will get an immediate answer. Some, though, will wait in queue for 3 to 5 minutes, assuming no overflow or other contingency. As a result, many in that set of callers would say that the service level is great, while a handful would describe it as poor.

- **Go with a “middle-of-the-road” objective, e.g., 80/20:** The objective 80/20 is fairly common because for many contact centers it is a reasonable balance between customers’ expectations and the practicality of having enough staff to meet the objective. It may or may not be right for you.
- **Relate to the competition – match or go higher/lower:** Benchmarking competitors or similar organizations, or using industry surveys that relay what others are doing can be used to help choose the center’s service level objective. Just keep in mind that the results reported by others and what they are actually achieving may be two very different things.
- **Adjust to minimize abandonment:** This approach asks, “How low can you go without losing callers?” This assumes that a higher level of service means lower abandonment and vice versa. A big flaw with this approach, however, is that abandonment levels and service levels do not always correlate. It is dangerous to assume that as long as callers don't abandon, service level

is acceptable. Abandonment will fluctuate as the seven factors of customer tolerance change. As a result, it is difficult to forecast, and choosing a service level around abandonment is building on the proverbial foundation of "shifting sand."

- **Use a combined approach:** Using an iterative process that combines the strengths of each of these methods is, overall, the best approach to choosing a service level. Consider at what level the center currently operates, run some calculations, consider what others are doing and (especially) the organization's unique brand, and assess what customers are saying and how they are reacting. New centers do not have the luxury of experience – whether good or bad – so starting a contact center from scratch will require more dependence on what others in the organization's industry are doing.

### 3.4 The Relationship of Service Level and Quality

#### Key Points

- **Service level and quality are complementary, and are not at odds.**
- **A poor service level contributes to poor quality as it puts workload pressure on agents and creates impatient and unhappy customers.**
- **Poor quality contributes to poor service levels by causing repeat contact, unnecessary service contacts, escalation of contacts, complaints to higher management, and other forms of waste and rework.**

#### Explanation

The conventional wisdom that quality and service level (or response time) are at odds and must therefore be “balanced” is one of the most fundamentally damaging misconceptions in the industry. Remember the two objectives that flow out of the definition of contact center management: Get the right people and supporting resources in the right places at the right times, and do the right things. They are complementary.

On the surface, it *does* appear that accessibility and quality are at odds. After all, you can have an excellent service level, but your agents can still:

- Misunderstand customers’ requests
- Enter the wrong information
- Relay the wrong information to callers
- Make customers mad
- Fail to accomplish the primary purpose (sell or service)
- Unnecessarily cause repeat contacts
- Miss opportunities to capture valuable feedback

But it is important to put this issue in context. Poor accessibility will rob the contact center of productivity. For example, as service level deteriorates, more and more customers are likely to verbalize their criticisms when their calls are finally answered. Agents will spend valuable time

apologizing to callers. Call handling time goes up and occupancy increases. If this condition continues, employee morale will sink. Turnover and burnout go up, as will recruitment and training costs.

When you consider the specific results of poor quality, the complementary relationship between accessibility and quality becomes even clearer. For example, what if data are not entered correctly? What if the caller doesn't have confidence the call was handled correctly? What if the agent didn't capture needed and useful information about the transaction? These problems contribute to repeat calls, escalation of calls, complaints to higher management, and callbacks, all of which further drive down service level. Or, if a customer doesn't receive a reply to an email as quickly as expected, he or she may send another; this can be the start of a similar cycle.

In short, there is no such thing as quality versus service level or response time. These objectives work hand-in-hand.