Module 8 – Real-Time Management

8.1 Interpreting Real-Time Information

Key Points

- **Random call arrival makes real-time management challenging, but important.**
- **Monitoring and responding to real-time reports should generally flow as follows:**
  - **Assess number of calls in queue**
  - **Assess longest current wait (oldest call)**
  - **Assess service level/average speed of answer**
  - **Assess agent status**
- **For outbound contacts, the primary reports that real-time analysts should consider are abandonment rate, which measures the number of callers who abandon before being connected to an agent, and agent status.**
- **For blended inbound/outbound call centers real-time analysts need to be aware of both inbound accessibility metrics as well as outbound pacing and abandonment rates.**

Explanation

The key to effective real-time management is to react appropriately to evolving conditions. Consequently, it is important to monitor developments and identify trends as early as possible. Random call arrival means that, at times, it will *appear* as though the center is falling behind even though it is staffed appropriately. But if there is a genuine trend, quick action needs to be taken to prevent further degradation of service.

Interpreting Reports

Service level is "rolling" history. The ACD has to look back at least some amount of time or at some number of calls to make the calculation. Consequently, even though service level is a primary focus in planning, it is not a sensitive real-time report. (In the strictest definition, it's not a real-time report.)

With many ACDs, management can define how far back the system looks to provide real-time service level status. The timeframe needs to include enough of a sample that reports aren't jumpy, but it also needs to be recent enough to be valuable. Also note that "screen refresh" does not correlate to the timeframe used for calculations.

Monitors may display updated information every few seconds, but that has nothing to do with how much data the ACD uses for the calculations that require rolling history.

Service level will reveal what has already happened, given recent patterns of unique call volume, random arrival, average handling time and staff availability. But it's important to realize that what is being reported is not necessarily an indication of what is about to happen.

On the other hand, the number of calls presently in queue is a real-time report, as is longest current wait and current agent status. Understanding the distinction between reports that are genuinely real-time and those that must incorporate some history explains apparent contradictions.

For example, service level may indicate 65 percent of calls are answered in 20 seconds, even though there are no calls in queue at the moment. Keep watching the monitor, though, and service level will begin to climb. Alternatively, service level may look high at the moment, even though an enormous volume of calls recently entered the queue. Give it a few minutes and, unless circumstances change, it will fall to the bottom of the barrel.

There will be at least several minutes delay before service level reflects the magnitude of a trend. As a result, for service level to have meaning, it must be interpreted in light of the recent past, calls in queue and current longest wait. If service level is the only measure considered, you could misread the situation.

Since the number of calls in queue foretells where service level is about to go (unless conditions change), it should be a primary focus, along with longest current wait. As circumstances dictate, you would then assess the state agents are in — signed off, auxiliary, handling calls, etc. — and make appropriate adjustments.

In sum, focus on reports in this order when managing in real-time:

1. Number of calls in queue: This is the real-time report most sensitive to changes and trends. Look at this first.

2. Longest current wait (oldest call): This is a real-time report, but behaves like a historical report (e.g., many calls can come into the queue, but longest current wait will take some time to reflect the problem). This report gives context to number of calls in queue. For example, if there are far more calls in queue than normal, but longest current wait is modest, the center is at the beginning of a downward trend. Now is the time to react.

3. Service level, average speed of answer, average time to abandonment and other measures of the queue and caller behavior: These reports of rolling history provide additional context to number of calls in queue and longest current wait. For example, if service level is low, but there are few or no calls in queue, then the problem is clearing and service level will begin to climb.

4. **Agent status**: This real-time report indicates how many agents are available and what modes they are in. Some managers suggest that agent status should be at the
top of the list. Their argument is that if agents are where they need to be, there won't be much of a queue in the first place. However, agent status can be difficult to interpret unless the condition of the queue is known. So what if few agents are taking calls, if few calls are coming in? In that case, it is appropriate for agents to be working on other tasks. With the right training on what real-time information means and the activity it is reporting, experienced agents and supervisors can scan and decipher these reports quickly.

Real-Time Reports for Outbound and Blended Centers

Real-time analysts in outbound and blended centers must monitor outbound measures. The primary outbound report that real-time analysts should consider is abandonment rate, which measures the number of callers who abandon before being connected to an agent. Many countries regulate outbound and blended centers based on average daily or monthly abandonment rates, so contact centers are under a regulatory obligation to keep abandonment at acceptable levels.

Agent status is also an important real-time report for outbound and blended centers. In order to handle the forecasted workload that was used for the basis of the schedules, the agents must be available for outbound calls, just as they do for inbound calls. If abandonment rate for outbound calls start to climb, analysts should first check agent status.

An important consideration when dialing in the predictive mode is that manual changes to the dialer, such as the addition or deletion of agents in a significant number or changes to campaign treatment options, will take at least twenty minutes to effectively register with the dialer. During those twenty minutes the dialer's software and algorithms will adjust to the new variables and the prediction will improve. A common mistake is to make too many changes in too short a period of time, not allowing the dialer sufficient time to adjust.

Because of the time it takes for the dialer to adjust to manual changes, real-time analysts in blended

centers must be especially careful not to react too soon. In most cases, spikes in call volume that are only one half-hour interval in length should not result in a change to the pacing of the outbound dialer. Rather, analysts should wait for a trend of three or more intervals where inbound accessibility metrics are below target before making changes to the dialer.

Real-time management in a blended center is particularly challenging due to the complexities of managing two sets of metrics, those for inbound call answering and outbound abandonment rates. Thus, blended centers should take care to develop an experienced team of individuals to manage their real-time management efforts.

Outbound/Blended Real-Time Management

While the argument can be made for inbound call centers that measuring "call abandonment" may not be an optimum metric to measure the success of an operation (due to the concept that caller tolerance is major factor in the decision to abandon), the same cannot be said for outbound dialing. With inbound call centers, it is fact that calls will bunch up due to random and peaked arrival. This will have a direct impact on abandonment rates and the ability to achieve answering thresholds. However, with outbound dialing, the "arrival" or "calls offered" is under the control of the dialer operator. Based on the speed or "pacing" of the dialer or the dialing pacing mode that is chosen, the outbound call center is able to proactively manage its abandonment rate. It is for this reason that a constant monitoring of the percent of calls abandoned within an interval, throughout the day and monthly, is an appropriate means to measure client experience as they interact with the outbound call center.

"Pacing" of a dialer is a strategy used to instruct the dialer when to offer calls to an agent and how to manage a list. Pacing simply put is the speed at which the dialer over dials in an attempt to match an agent to a caller. It is typically managed through abandonment levels or caller wait times prior to being offered an agent. The mode of pacing that call center chooses as a strategy directly impacts abandonment. The strategic intent of the outbound center, in most cases, will make the choice for the mode of pacing obvious. It is important to note that some centers will employ one or more of the pacing methods at one time, depending on treatment strategy for the campaign.

Modes of Outbound Pacing

- Predictive – The dialer will "over dial" available numbers in an attempt to offer a call to the agent when it predicts they will be available
- Progressive – The dialer will begin to call all available numbers for the next record once the previous call has ended
- Preview – The agent will direct the dialer to what numbers to call and when

Reacting too soon and making too many changes in a short period of time is one of the most common pitfalls blended call centers make when managing a peak in calls. One of the "immutable" laws of using a dialer is that it takes roughly twenty minutes, regardless of brand of dialer, for predictive pacing to adjust to the addition/deletion of agents or a change to pacing speed. For this reason, when there is an inbound spike, it will take twenty minutes for your dialer to adjust to having fewer agents available for offered calls and your abandonment will also spike if you take too many agents away at once.

Written for ICMI by Rob Archambault.

8.2 Alternatives for Providing Real-Time Information

Key Points

- **Agents should have access to information on the queue and be trained to interpret it correctly.**
- **There are a number of alternatives for providing real-time information, including:**
    - **Agent workstations**
    - **Agent telephone sets**
    - **Wall- or ceiling-mounted readerboards**
    - **Supervisor monitors**
    - **Low-tech alternatives, e.g., updated white boards**

Explanation

Everyone in your contact center needs to be aware of the impact each agent has on the queue. The message, as it relates to real-time management, is clear: When the queue is backed up, each person makes a big difference.

This issue sheds light on the importance of training agents on how a queue behaves (e.g., how fast it can spin out of control) and, with cautions discussed below, providing them with real-time information so they can adjust priorities as necessary. Real-time information can be delivered via:

- Windows with queue information on desktop displays
- Supervisor monitors
- Wall- or ceiling-mounted readerboards
- Displays on telephones programmed to give queue statistics
- Mobile apps
- Low-tech alternatives, e.g., regularly updated results on easels or white boards throughout the center (not an ideal solution, but better than nothing)

Queue information must be complemented with appropriate training so that agents know what to look for and how to react. There are two things directly within the control of agents: being in the right

place at the right times, and doing the right things (schedule adherence and quality, in industry terms). A backed-up queue does not mean you should change the way you handle work, the process necessary for handling each contact with quality. There may be times when it is appropriate to turn off information when all it can do is add stress to a stressful situation. Real-time queue information must be interpreted in that context.

8.3 Setting Real-Time Thresholds

Key Points

- **ACDs, desktops and wall display systems allow you to establish a variety of display thresholds.**
- **Three thresholds are recommended that indicate various levels of real-time response based on how many calls are in queue versus how many are expected to be in queue.**

Explanation

Various priority thresholds can be established in some ACD and wall display systems. For example, when the queue begins to back up, the information can be color-coded yellow. When it is in bad shape, it is color-coded red.

The problem is, the thresholds are often set arbitrarily. Further, agents often do not understand what is expected of them at different levels. If that is the case, real-time information will raise everybody's stress level. And agents might feel like it's their fault that they can't clear up the queue. Proper programming and training are necessary.

Setting Thresholds

It is recommended that the thresholds be set as follows:

- **First threshold**: Generally the first threshold should be set for one call in queue. Agents should proceed normally, and no tactical adjustments are required.

- **Second threshold**: The second threshold should indicate that there are more calls in queue than the average expected for the desired service level. The average expected for the desired service level can be found in the Q2 column of an Erlang C calculator. Routine adjustments should be made (e.g., postpone flexible work) to get the calls answered.

- **Third Threshold**: The next threshold should indicate that there are more calls in queue than the agents can handle. In this case, more involved real-time tactics are required (e.g., calling in reinforcements).

Some systems can be programmed to adjust thresholds as calling loads change (10 calls in queue may be no problem during a fully staffed shift, but would be a big problem for two people handling calls at 3 a.m.).

Overflow Thresholds

Service can often be improved by changing call-routing thresholds between groups or sites. A common strategy is to overflow to agents who are assigned to work that is not as time-sensitive as service level contacts.

Most modern ACDs are capable of if-then programming logic to automate this process. But there are cases that may require some adjustments. There are several things to keep in mind when setting overflow thresholds:

- Secondary or tertiary groups must be sufficiently equipped and trained to handle the contacts.
- Management must determine in advance that calls to the primary group should take precedence over the work secondary groups otherwise would be handling.
- Thresholds should reflect these tradeoffs, and should be established with an understanding of how queues behave.

Blended (inbound/outbound) centers can design overflow thresholds that reassign outbound agents to the inbound queue when service level is suffering. However, everyone in blended centers must know the strategic intent of the center and whether outbound calls or inbound calls receive a higher priority. Most often, centers place a higher priority of servicing inbound calls since the center cannot control when those calls arrive. However, a center whose strategic intent hinges on successful outbound call campaigns may choose not to overflow even when service levels in the inbound queue are low.

8.4 Informational and Delay Announcements

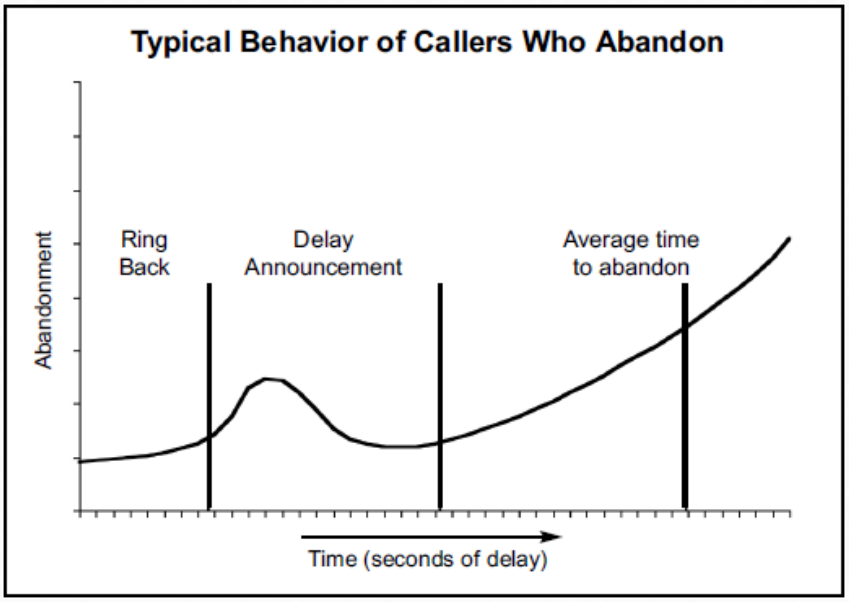
Key Points

- **Delay announcements are adjustable, and should reflect real-time circumstances.**
- **The first delay announcement recognizes callers, explains the delay and promises that the calls will be answered.**
- **The second delay announcement is designed to give callers assurance that they haven't been forgotten.**
- **Careful consideration should be made about what is said in delay messages and how frequently a customer hears them.**

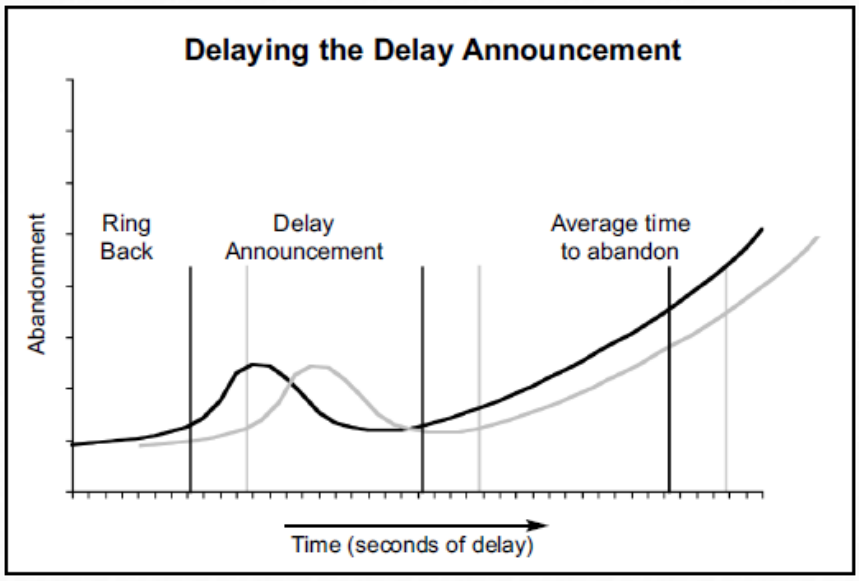
Explanation

Most call centers provide announcements to callers who wait in queue. The first announcement recognizes callers, provides reassurance they are in the right place, and promises that the calls will be answered. This announcement can also advise callers of what to have ready for the call (e.g., account number), and provide alternative contact methods (e.g., "visit our Web site at www..."). It is typically provided to the caller prior to sending the call to the queue. This means that a caller who decides not to enter the queue is not counted as an abandoned call and that it does not reduce service level for the interval. For this reason, it is important that attention is paid to the number of callers who choose to disconnect before queuing.

The typical behavior of callers who abandon can provide insight into the use of delay announcements. Callers who hang up when they hear the first delay announcement are called "fast clear-downs." The customer may have dialed the wrong number or may just have changed their mind and decided to bail out when they didn't get right through to an agent.



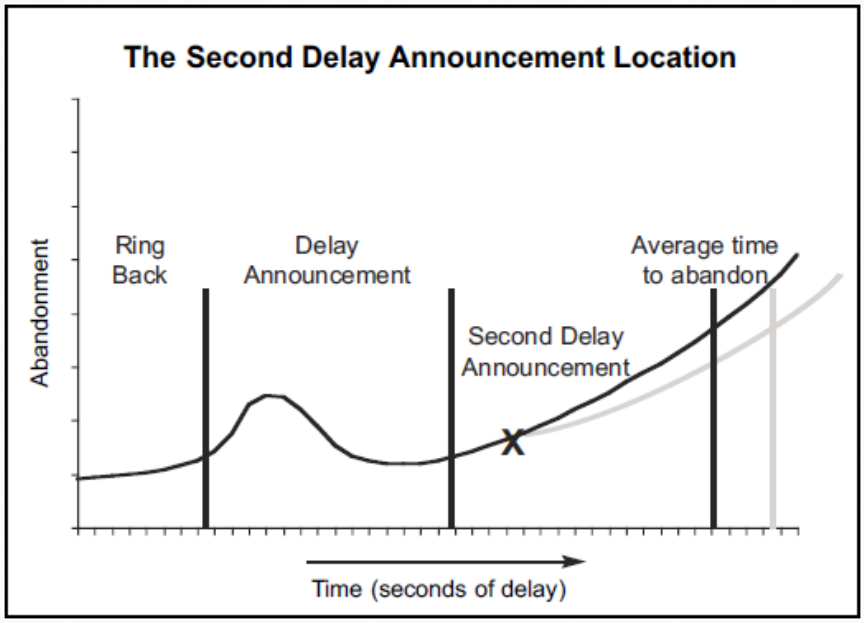
Some centers have found that repositioning the first delay announcement can lower abandonment. For example, if the delay announcement is normally set to come on just after a caller enters a queue, moving the threshold out to provide additional rings can buy the agent group additional seconds to get to callers before they become fast clear-downs. Further, because callers don't always mentally register that they are in a queue until they hear the announcement, they may wait longer.



An important consideration when delaying the delay announcement is that this technique will

actually increase average speed of answer and reduce service level. What will be achieved is the higher priority of getting to as many callers as possible before they give up and hang up.

Abandonment may also be reduced by adjusting the position of a second announcement. For example, if average time to abandonment is 50 seconds and the second delay announcement is set for 60 seconds, adjusting the message to play earlier may cause more callers to wait longer. Remember the purpose of the second delay announcement is to give callers who are about to abandon renewed hope that their call will be answered: "We apologize for the delay; thank you for continuing to hold ..."



The first and second delay announcements are valuable, but good judgment is needed when using announcements. Careful consideration should be made about what is said in delay messages and how frequently a customer hears them.

- If you always have long delays, eliminate the “we are experiencing higher than normal call volumes” from the message. Customers don’t want to hear it unless it is truly an unusual circumstance.

- Change the messages so that there is variety in what a caller hears. Provide information about what is new, how to use other forms of support, best times to call back, etc. in order to keep the customer engaged.

- Repeating announcements of the same message tends to make things worse. Research shows that customers dislike hearing the same message time and time again, so it is worth the effort to vary these messages.

- Delay messages should reflect the branding style of the products and services offered by the contact center. An example of this is the delay messaging used by Southwest Airline. While waiting in Southwest Airline's queue, customers listen to contact center agents telling jokes. This would be inappropriate for some organizations, but aligns with the Southwest Airline brand. Be sure the content, tone and timing of delay messages match the brand of the organization.

Outbound centers may be required to play delay messages to customers who are not connected to agents within a certain number of seconds. Check with legal counsel to determine if any such regulations apply to your center.

8.5 Establishing and Using Real-Time Tactics

Key Points

- **Establishing an effective real-time escalation plan involves:**
  - **Identifying feasible real-time tactics**
  - **Determining the conditions for which each should be implemented**
  - **Monitoring conditions**
  - **Deciding on adjustments necessary**
  - **Coordinating and communicating changes to all involved**
  - **Implementing the tactics**
  - **Assessing how well the escalation plan worked**

- **Most contact centers used a tiered approach to categorize appropriate tactics based on the severity of the situation.**

Explanation

To achieve service level and response time objectives in real-time, the center will need to make appropriate tactical adjustments as conditions change. An important principle in effective real-time management is to outline a workable escalation plan that is understood *before* a crisis occurs.

Establishing an effective escalation plan involves:

- Identifying feasible real-time tactics (ahead of time)
- Determining the conditions in which each should be implemented (ahead of time)
- Monitoring conditions (real time)
- Deciding on adjustments necessary (real time)
- Coordinating and communicating changes to all involved (real time)

- Implementing the tactics (real time)
- Assessing how well the escalation plan worked (after the fact)

When identifying feasible real-time tactics, most contact centers use a tiered approach.

Level 1 Tactics

The first level of action involves routine, commonsense adjustments.

- **Focus on agent status**: Many use a variation of the time-honored phrase: "Everybody take a call!" This is generally directed toward people on the floor who are not currently handling calls. It also can be for agents stuck in wrap-up mode.

- **Postpone flexible work**: At this level, agents make routine adjustments to work priorities. Flexible tasks are postponed. If agents are handling contacts that are not as time-sensitive and can wait — social interactions like updating the company Facebook posts, email, mail, outbound calls, or data entry — those agents can be temporarily assigned to the calls in queue. Blended centers should have guidelines for real-time analysts that explain the strategic priorities of their center. Some blended centers may be willing to sacrifice inbound service levels for a specific outbound campaign because of its strategic impact. Real-time analysts must understand the strategic intent of the contact center so that they can make appropriate decisions as they respond to developments in the queue.

A note of caution: Make sure that agents understand that speeding up their rate of speech will not help. Callers can usually sense they are rushed, and will often dig in their heels to slow things down. However, agents shouldn't go beyond what is necessary to completely satisfy the caller's stated objectives and handle the call with quality. There's a line somewhere, and commonsense applies.

Level 2 and Beyond

If the workload still outpaces the staff required to handle it, the contact center can move on to more involved real-time alternatives:

- **Reassign agents to groups that need help:** Depending on the status of the queue for other agent groups, some agents from other groups may be able to be reassigned temporarily to the group that needs help. Of course, only agents who have been trained to handle the contacts in that queue can be reassigned.

- **Situation-specific messages**: Another possible Level 2 activity is to change system announcements so that they offload what would otherwise be routine calls. Utilities use messages such as, "We are aware of the power outage in the Bay Ridge area, caused by nearby construction. We hope to have power restored by 11 a.m. We apologize for the inconvenience. If you need further assistance, please stay on the line; one of our representatives will be with you momentarily."

- **Redirection messages**: More routinely, calls can be directed elsewhere: "Thank you for calling ABC airline. If you would like to use our automated flight arrival and departure system, please say or press ..." Some centers also give callers the ability to check the status of an order, listen to specific product information, or hear answers to commonly asked questions while they wait and without losing their place in the queue.

- **Adjust call routing priorities:** Changing call-routing thresholds between groups or sites may also help to improve circumstances. Most of today's routing systems are based on a form of "if-then" programming logic to automate this process. But there are cases that may require some adjustments. And if the network sends fixed percentages of calls to various sites, those thresholds may need to be adjusted. Blended centers may also have to

- **Triage contacts**: Another tactic is to triage contacts in a way that makes sense. For example, the most important contacts can receive priority over contacts that are less urgent or less valuable to the organization. Prioritizing by call type can be based on numbers dialed, routing selections, customer identification and other criteria.

- **Use supervisors to help handle calls**: It also may make sense for supervisors and managers to help handle customer contacts. However, this approach must be well thought-out, because if they are unavailable when agents need help, the situation could further deteriorate. Some union agreements restrict supervisors and managers from handling contacts, but if allowed, this can be an effective tactic.

- **Take messages for callback**: Some centers take messages for later callback, a capability that is greatly facilitated by virtual queue technologies. However, this approach does not work well in all cases. Potential challenges include: How do you ensure that the callbacks are timely if you're busy now? What is your policy when you reach the caller's voicemail? You may have to experiment to find out whether it's workable in your environment.

- **Mobilize the swat team and others**: Other Level 2 tactics include calling in a swat team, bringing in agents who are on reserve, routing some calls to established outsourcers,

adjusting the placement of delay announcements and generating controlled busy signals.

Post-Analysis Improvement Process

On the other end of the crunch (or unexpected slack) is an important but sometimes neglected aspect of real-time management: analyzing what happened so that you can prevent recurring problems. How well did the escalation plan work? Were the right tactics deployed? What can be done differently? This analysis will help fine-tune the escalation plan and improve the planning process. It is especially important if real-time tactics are being used to help with more than two or three significant crises per week.

Real-Time Recovery Planning

An effective real-time queue management program is essential to running an efficient inbound contact center, but it's often the piece that's left out of the planning process. Let's look at a few considerations for establishing or updating a real-time recovery program.

Continually Update the Plan

Your ability to institute a successful real-time recovery program starts with an effective planning and scheduling process. This process must include the ability to look ahead to the coming week and identify the intervals that lack the minimum number of phone agents required to meet your service level objective. This capability is a common component of most workforce management systems, or it can be manually tracked via a spreadsheet or database application. If you do not have a process in place that allows you to look ahead and review staffing gaps by interval, you should implement one before moving forward with a real-time recovery plan.

Once you have a process that allows for ongoing staffing gap analysis by interval, it is important to keep it updated. If you create schedules several weeks in advance, they need to be continuously updated with all changes that will affect the number of employees planned for incoming calls. This includes changes to the volume forecasts, last-minute agent training sessions and/or meetings, short-term disability leave, etc.

Last, and most important, your plan must be updated with the last-minute changes first thing in the morning (e.g., sick leave, broken-down cars, sick children, etc.). This will give you an accurate picture of the workforce availability and will provide you with ample time to review alternatives for any intervals that look hopeless. Keep in mind that your current day planning does not end after your initial morning update. Additional unexpected events will influence your workforce throughout the day and the plan must be adjusted accordingly.

(continued…)

Communicate Expectations to the Front Line

The key to a successful real-time recovery program is the communication of expectations. The first step is to develop a process of communicating the expected workforce variances and any last-minute changes to the plan. This can be accomplished by consolidating the expected workforce variances by interval for an entire week on one spreadsheet. The spreadsheet, along with the ongoing updates, could then be emailed to your staff or posted on an intranet site.

Providing this continual "snapshot" of the workforce and workload distribution by interval will eliminate many of the queue surprises that tend to catch everyone off guard. If the snapshot shows fewer people staffed than needed to meet the minimum service level, the odds are you're going to have calls in queue and everyone should be aware and prepared.

Once the plan is communicated to the front line, expectations should be clear as to what actions are to be taken. For instance, if you have cross-trained agents who can handle response time activities (e.g., email, fax, etc), you may have them log onto the phones once your queue threshold has been exceeded. Better yet, in the intervals when the plan is at a significant deficit, you could have them log onto the phones in advance, which will help to avoid the painfully long process of driving down the queue. You'll need to recover the time lost against the response time activities and move phone agents into non-inbound modes during intervals when the plan illustrates excess capacity.

Don't Set a Reaction Based on a Static Number

Once you've defined and communicated the actions to be taken, you will need to determine when the program should be implemented and when to escalate to the next level. Using static indicators as criteria for implementation will result in over- or under-reacting in many cases.

For example, let's say the first phase of your plan is to begin when your expected interval-staffing shortfall is at negative five. A staffing deficit of negative five will result in significantly longer hold times when the required staffing is 20 than it will when the required staffing is 40. Using a single number as your threshold tends to mask the urgency in your lower volume intervals. A good method for setting your proactive adjustment threshold is to use a percentage approach — plan to invoke different phases based on the deficit percentage and not a static number.

This same approach should be used for those intervals when the calls don't arrive as planned and you need immediate help. You'll first have to work with an Erlang program to get a feel for the "planned" number of calls in queue based on the expected volume for the time of day. Next, determine how long the threshold can be exceeded before enacting the plan. It'll take a few attempts to get this right but, once established, it will definitely reduce the number of "hair-on-fire" events.

Written for ICMI by Tim Montgomery.