2007

Improving Call Center Operations Using Performance-Based Routing Strategies

J. Stanley

R. Saltzman

Vijay Mehrotra
University of San Francisco, vmehrotra@usfca.edu

IMPROVING CALL CENTER OPERATIONS USING PERFORMANCE-BASED ROUTING STRATEGIES

Julie Stanley, Robert Saltzman and Vijay Mehrotra

San Francisco State University

ABSTRACT

The article presents a simulation study of performance-based call routing strategies using a variety of routing rules based on historic data such as average handling time and first call resolution rate. We demonstrate the relative benefits of various performance-based call routing strategies using actual data from a financial call center. In particular, our modeling results indicate that call routing based on adjusted average handle time (AAHT) and z-scores of AAHT are two strategies that can substantially improve overall call center performance and customer satisfaction.

I. PROBLEM DESCRIPTION AND SYSTEM BACKGROUND

"Where should we route that call?" and "Who should take that call?" are two key questions often asked in service based call centers. Finding good answers to these questions can help call centers achieve much better performance. In fact, rules for routing incoming calls have a great impact on both call center performance and customer satisfaction. When customers have to wait too long to reach an agent or do not have their calls handled successfully by the first agent they encounter, their satisfaction and good will decrease (Levin, 2007). Richard (2002, p. 5) finds that customer dissatisfaction increases exponentially with each poor experience, and often results in lost business.

Three often-considered measures of call center performance are average speed to answer (ASA), average handling time (AHT), and first call resolution (FCR) rate. Many heuristics and rules have been studied and employed, with first-in, first-out (FIFO) rules being the simplest to implement. However, FIFO often produces situations where a call is not handled by the most qualified agent, leading to high average handling times and low first call resolution rates.

Skill-based routing (SBR) has received attention recently in industry and academia alike (IEX, 2003). SBR routes calls based on a matrix of static ratings assigned to each agent derived from their skill in handling each call type (Wallace and Saltzman, 2005). While SBR improves overall performance, it also has some drawbacks, including unnecessary real or virtual team divisions based on these skill sets that drive down efficiency (Richard, 2002, p. 243). In addition, maintaining skill sets rating data to use with SBR algorithms requires time and resources.

More recently, Sisselman and Whitt (2007) present a mathematical program to maximize the total value derived from n different agents handling m different call types, where each agent j is assumed to have a specific value $v_{i,j}$ for handling a call of type i. Each value $v_{i,j}$ is a real number that may correspond to financial value, to likelihood of first-call resolution, or to preferences of individual agents for handling particular types of calls. While maximizing overall system value, they seek a call routing policy that meets traditional constraints involving customer waiting time metrics. This article, by contrast, uses simulation to accurately incorporate the impact of random call arrivals, handle times, and call resolution on performance.

New heuristics, called performance-based call routing (PBR) rules, are now being studied. The focus here is to determine which PBR rules are most effective through the use of simulation in the Arena software environment (Arena, 2005; Kelton, *et al.,* 2007). These rules are evaluated based on a comparison to a baseline model using FIFO routing. Like SBR rules, PBR routes calls based on actual past performance of agents with calls over time. However, PBR differs in that the matrix for rules is derived from actual performance measures for each agent by call type rather than ratings. A number of performance measures have been used to design PBR rules, including AHT, FCR rate,

waiting time (W), and queue lengths (L), with the goal of maximizing the throughput of successfully resolved customers while ensuring that no customer types wait too long.

We studied a financial call center, a stochastic system having *M* call types and *N* agent groups (see Figure 1), and assumed that all agents were cross-trained and able, although not always the best, to handle all call types. In practice, agents are heterogeneous because their skills actually do vary. Arrival rates and processing times vary by call type as well. The two central questions that must be addressed by a system with PBR are: (1) When a call comes in and finds agents from more than one agent group available to handle it, which agent should handle the call? (2) When an agent from one group finishes a call and finds more than one type of call waiting, which call should it choose?

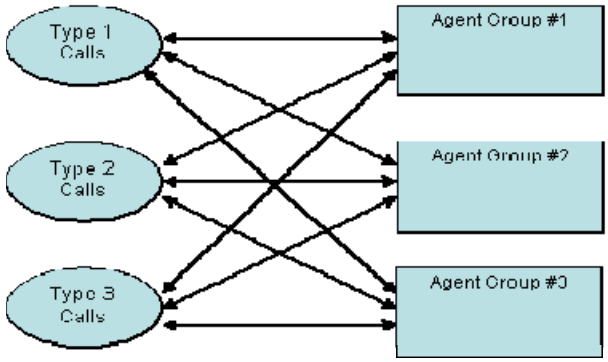

Figure 1. Simple Representation of System (with *M* = *N* = 3).

There are many ways to route calls in this type of system. Some routing rules are *static*, meaning that the same agent always handles a particular call type, while others are *dynamic*, with calls getting assigned to agents based on the current condition or state of the system. We focused on dynamic call routing rules, and used FIFO as the baseline of comparison with various PBR strategies. Under FIFO, incoming calls are held in a single processing queue from which the first call is taken into processing whenever an agent becomes available, i.e., agents and calls are paired without regard to past performance.

In PBR, both key questions regarding call-agent pairing must be carefully addressed to improve call center performance. Past performance is accounted for by the rules assigning agents to incoming calls and assigning calls to free agents. For example, in the AAHT experiments described more thoroughly later, an incoming call is sent to the free agent whose AAHT value for that call type has been the lowest historically. Similarly, agents coming free choose the next waiting call type to serve based on whose historic AAHT value has been lowest for them.

While this article employs simulation to study how call centers should route incoming calls, a wide range of articles have been published on call center management, employee issues, and customer concerns (e.g., see the survey by Gans, *et al.* 2003), a few of which we now review to provide more context for our study. Armistead, *et al.* (2002) study managerial issues through focused case studies and interviews with operations managers, finding that call center employees exhibit a growing professionalism, in keeping with the increasingly prominent role of call centers in market value chains. Managers, meanwhile, face difficult tradeoffs between customer service and efficiency, making it critical to choose technology, roles, skills, and competencies carefully. Houlihan (2002) also notes the tensions that call center managers face between costs and quality, flexibility and standardization, and constraining and enabling agent job design. In studying four actual call centers, she identifies managers who combine different degrees of commitment to their agents with various levels of agent empowerment; however, it's unclear how these managerial modes specifically affect the customer experience and call center performance.

Rather than studying measures of internal efficiency and control, Dean (2004) looks at customer *expectations* of call center service quality. From a survey of an insurance provider's customers, she finds that customers have very high minimum (adequate) expectations that are independent of their predicted expectations. She concludes that customer satisfaction may be harder to achieve than expected, but can be influenced by paying attention to the features that comprise customer orientation. Later, Dean (2007) studies the impact of *perceived* agent service quality and customer focus on customer loyalty and emotional attachment to the company, rather than examining and making use of actual service quality by agents, as we do here. Through survey data from online banking and insurance businesses, she finds various degrees of correlation between the variables studied, but offers no concrete, actionable advice for improving the call center's actual performance.

II. INPUT DATA AND MODEL ASSUMPTIONS

Data for this project come from a Fortune 50 financial firm's call centers, with FCR rates and AHTs being key inputs. For the site experiments, data were drawn from five call types (M = 5) at four agent group call centers (N = 4). The five call types were 1) Inquiry, 2) Overdraft Protection, 3) Payment, 4) Reinstate-ment, and 5) Web Related, while the four agent groups or sites were 1) Columbus, 2) Mumbai, 3) San Antonio, and 4) Surrey. For the team experiments, data were drawn from five call types (M = 5) and five agent group teams (N = 5) identified by supervisor. The five call types were 1) Account Maintenance, 2) Authorization Security, 3) Card Replacement, 4) Credit Related, and 5) Inquiry, while the agents teams were 1) Campbell, 2) Ebanks, 3) Lee, 4) Schlientz, and 5) Taft. Each model (except the AAHT+Extra Agent model) had two agents per agent group.

Simulating a particular institution's call centers was not as important to us as finding rules that would broadly apply to a variety of call center environments. As incoming calls to a call center usually follow a Poisson arrival pattern, an exponential distribution with a mean of 0.48 minutes was employed to simulate call interarrival times and create a utilization rate of 85-95% across agent groups. For call volumes by call type, a discrete distribution was used in which each call type represented 20% of the incoming call volume.

To simulate processing time, an exponential distribution was used with mean AHT in seconds (left side of Table 1) provided by the financial institution call center. In addition, mean FCR rates provided by the same (right side of Table 1) were used to decide whether or not a call was successfully resolved on first call.

AHT - Team Experiments

Call Type	Campbell	Ebanks	Lee	Schlientz	Taft
Account Maintenance	223.6	164.7	216.3	219.4	150.1
Authorization/Security	190.0	191.3	191.0	202.1	173.2
Card Replacement	224.1	127.9	180.3	213.3	118.0
Credit Related	238.0	221.4	202.6	290.1	181.4
Inquiry	173.3	132.5	143.6	186.1	116.1

FCR Rate - Team Experiments

Call Type	Campbell	Ebanks	Lee	Schlientz	Taft
Account Maintenance	0.9569	0.9537	0.9530	0.9582	0.9495
Authorization/Security	0.9370	0.9263	0.9437	0.9395	0.9307
Card Replacement	0.9445	0.9438	0.9392	0.9447	0.9364
Credit Related	0.9663	0.9506	0.9730	0.9710	0.9549
Inquiry	0.9310	0.9265	0.9339	0.9323	0.9248

Table 1. AHT and FCR rates by call type and agent group (team experiments).

For search rules assigning calls to agents and vice versa, a number of inputs are used in the seven PBR models. For instance, Adjusted Average Handling Time (AAHT) measures both speed (AHT) and accuracy (FCR) and is defined as the ratio AHT/FCR (see Table 2). These data are inputs to the AAHT models that try to find the best agent to assign to a given call when many agents are available, and the best call to assign to a newly freed agent when many calls are on hold.

AHT/FCR - Team Experiments

Call Type	Campbell	Ebanks	Lee	Schlientz	Taft
Account Maintenance	233.7	172.7	227.0	229.0	158.1
Authorization/Security	202.8	206.5	202.3	215.1	186.1
Card Replacement	237.2	135.5	192.0	225.8	126.0
Credit Related	246.3	232.9	208.2	298.8	190.0
Inquiry	186.1	143.0	153.8	199.7	125.5

Table 2. AAHT values for by call type and agent group (team experiments).

Z-scores (or normalized scores) of AAHT values (see Table 3 below) were also used as inputs for search rules built into the Z-Score of AAHT models. The intent of using z-scores was that they might better indicate the agents' proficiency relative to other agents. Usage of z-scores is described further in Section IV.

Team Experiments
Zagents - Z-score by call type relative to all calls across all agents

Call Type	Campbell	Ebanks	Lee	Schlientz	Taft
Account Maintenance	0.827	-0.878	0.640	0.695	-1.285
Authorization/Security	0.023	0.374	-0.023	1.190	-1.563
Card Replacement	1.060	-0.939	0.171	0.835	-1.126
Credit Related	0.266	-0.057	-0.649	1.525	-1.085
Inquiry	0.800	-0.607	-0.255	1.241	-1.178

Team Experiments
Zcalls - Z-score by agent type relative to all agents across all call types

Call Type	Campbell	Ebanks	Lee	Schlientz	Taft
Account Maintenance	0.487	-0.131	1.118	-0.123	0.031
Authorization/Security	-0.721	0.685	0.209	-0.486	0.928
Card Replacement	0.626	-1.027	-0.171	-0.206	-0.999
Credit Related	0.982	1.320	0.425	1.706	1.054
Inquiry	-1.374	-0.846	-1.580	-0.891	-1.015

Table 3. Z-scores of AAHT values by call type and agent group.

To keep the simulation models and analysis manageable, we made several simplifying assumptions. First, all agents were assumed to be cross-trained, yet heterogeneous. Second, the number of call types, agent groups, and agents per group were chosen so as to create a good distribution of all call types that could be handled by all agent groups. Third, the subset of call types chosen from the financial institution data were selected so as to ensure a set of calls that would be spread across as many agent groups as possible when using search rules to assign best agents to calls and vice versa. Fourth, a value of one minute was (arbitrarily) chosen for purposes of simulating a delay prior to a customer calling back when a call was not resolved. Fifth, testing was based upon steady state conditions devoid of any seasonality regarding time of day, month, or year. A one hour warm-up period was used in all replications to achieve steady state conditions. Finally, caller abandonment was not included in our models in order to simplify both the programming and the analysis.

Output statistics collected for analysis from each experiment represent the means across 30 replications of a 12-hour day and include: (1) FCR Rate by call type and overall across call types; (2) AHT by call type and overall across call types; (3) ASA by call type and overall across call types; and (4) Total number of calls resolved. We focused on increasing average FCR rate, and decreasing average handling time and ASA, while ensuring there were no strongly negative impacts to overall number of calls resolved or individual call types. Finding rules that lead to positive output for all three performance measures was sometimes difficult.

III. MODELS TESTED

In the *base case* model, a FIFO strategy is employed to route calls through the call center, *i.e.,* calls are processed based on their arrival time, with the oldest call always getting processed first, regardless of what agent could best handle it. If a single agent is free when a call arrives, the call is immediately assigned to this agent. If multiple agents are free when a call arrives, the agent assigned to handle the call is chosen using a cyclical selection rule. When an arriving call finds all agents busy, it is queued at a call processing module; agents take calls from the front of this queue as they become free.

After being served, calls take one of two directions based upon what happens at the decision module named "*Resolved?"* (bottom of Figure 2). Calls taking the *True* (resolved) path pass through several modules that record the number of calls completed, time to complete the calls, the FCR rate, etc. The value of a call attribute named *CallsToResolve* is also evaluated to determine whether a call should be recorded as having been resolved on first try. Calls taking the *False* (unresolved) path are sent to a module that increments the call attribute *CallsToResolve* (which is set to one at call creation), so that the call is no longer counted as having been resolved successfully on first call. The call is then delayed briefly prior to being sent for additional processing. These calls are not included in the number of completed calls until they are successfully resolved.

PBR models seek to optimize the pairing of calls and agents so as to increase the achieved FCR rate, decrease AHT and/or decrease ASA. These models are more extensive than the FIFO model, with the most noticeable differences being in how agents are selected to handle incoming calls, how calls are selected by free agents, and the use of multiple call processing modules. Figure 2 shows the logical modules for the AAHT performance-based model. Call creation and assignment of call attributes are handled in the same manner as in the base case FIFO model. However, right after a call arrives agent groups are searched to determine whether there are any free agents to handle the call. If any group has a free agent, a search takes place to determine the best available agent to take the call. The search module scans an expression (e.g., *AAHT*) across all free agents for the specified call type in order to find the lowest (AAHT model) or highest (FCR Rate model) value for the call and agent pair. The call type is fixed since the incoming call type is known.

For incoming calls where no free agent is found, the call is routed to a queue (*Wait Station*), where it is held indefinitely until being chosen at some later point by a free agent. Incoming calls that find a free agent available to process it are routed to the processing module for the particular agent type assigned to the call. There is a processing module for each agent type used in the experiment. After processing is complete, the agent coming free is marked as such and a new search to find the best call for the agent to take from the queue takes place. In the AAHT model, expressions for all calls in the holding queue are searched to find the one that has the lowest value for the call-agent pair. Here, the agent type is fixed in this search since the agent type is known. The selected call is then removed from the holding queue and assigned to the free agent. The original call entity is routed to a decision module that determines whether not it was successfully resolved, as in the base case. If no calls are waiting when an agent comes free, the agent stays idle until a new incoming call is routed to it. The remaining components in the AAHT model are the same as in the base case.

IV. ANALYSIS OF ROUTING STRATEGIES

Numerous experiments were run in order to evaluate various PBR strategies. These can be organized into three different categories: 1) experiments with input data coming from consolidated *team* level data within a site, 2) experiments with input data coming from consolidated *site* level data, and 3) experiments with scaled up resources and call arrival rates based upon team level data. In each category, up to seven PBR strategies were tested in addition to FIFO. Due to limited space, this article only discusses results from the "team experiments" category using two of the PBR strategies. The complete set of experiments is reported in (Stanley, 2007).

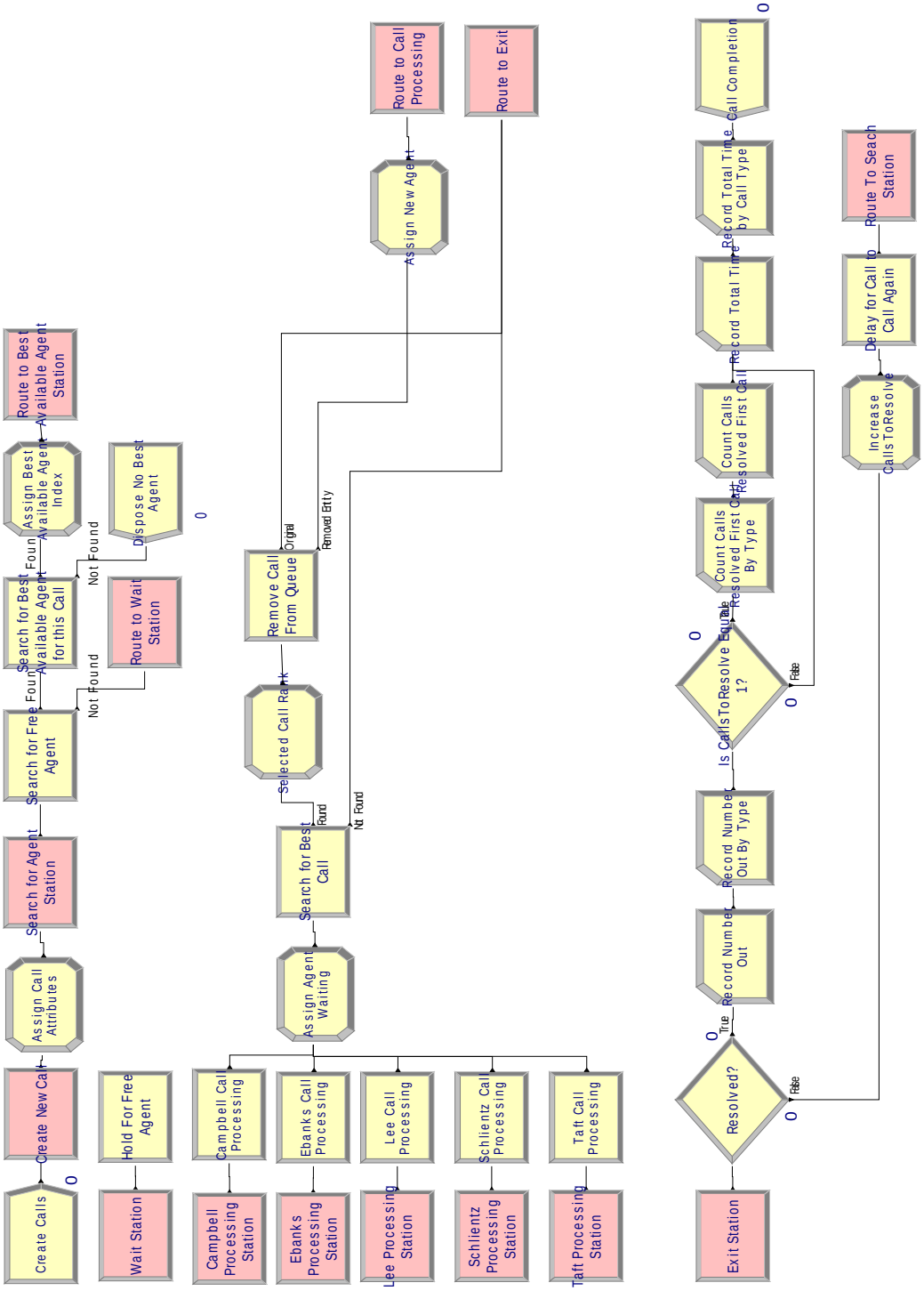

Figure 2. Arena flow modules for the AAHT model (team experiment).

The AAHT routing strategy balances speed and accuracy in call handling. The AAHT ratio captures the tradeoff between speed (AHT) and accuracy (FCR) in measuring overall performance of individual agents, teams, or sites within a call center. Low average handling time does not necessarily translate into high FCR rates ("How DHL", 2005). In fact, sometimes low AHTs actually indicate poor call handling quality (more unresolved calls) and lead to a decrease in FCR rate (Richard, 2002, p. 39). In experiments using this strategy, the AAHT value was used as an input into the search module identifying the best call-agent pairing.

Output from the team experiment showed a 0.21% increase in the FCR rate, a 3.71% decrease in AHT from an average of 194 down to 187 seconds, and a 50% decrease in overall ASA from an average of 183 to almost 92 seconds (all changes are relative to FIFO results). Using a labor rate of $30 per hour per agent and an estimated cost of $6/call, the savings for the call center and its customers can be calculated as follows. Raising the FCR rate saves 0.21% x 1895 calls/day x $6/call or $23.88/day in call handling. Reducing AHT saves an average of 3.71% x 1895 calls/day x 194 seconds/call = 13,639 seconds or 3.79 hours agent time/day, which translates into $113.70/day, for a total of $137.58 per day in direct savings. Moreover, lowering the ASA means a reduction in customer waiting time of 50% x 1895 calls/day x 183 seconds/call = 173,462 seconds or 48.18 hours customer time/day. Among all rules tested, the AAHT rule's direct savings were second only to the Z-score strategy, and its savings in customer time was the best.

Results by individual call type, however, are not uniformly positive. While most call types see large improvements in all performance measures, one call type suffers at the expense of these improvements. In particular, ASA for Credit Related calls increased by 65% from 180 to almost 298 seconds. Thus, to achieve overall performance improvements, some customer types would have to wait considerably longer, on average. While there are always tradeoffs in determining the best overall routing strategy to implement, much longer waits for 20% of a call center's calls might lead to greater call abandonment and customer dissatisfaction. Though our models do not incorporate call abandonment, it is clear that lost calls and ultimately lost customers are a real cost to business that ought to be taken into account.

The Z-score routing strategy, like the AAHT strategy, balances speed and accuracy in call handling, but also tries to ensure that agents' performance relative to other agents with each call type is considered. For example, to determine the call that an agent chooses upon becoming free, z-scores (normalized AAHT values) are examined for each call type. For example, in the top half of Table 3, *Zagents* are z-scores found across agent groups for each call type (i.e., the mean and standard deviations come from the values in each row). A newly freed agent chooses the call type for which this agent provides the largest benefit. For instance, if the free agent is on the Lee team, each waiting call is evaluated to find the one with the lowest value from the *Zagents* table for a Lee agent. Assuming all call types are waiting on hold, a Lee team agent selects the first waiting Credit Related call. Similarly, arriving calls finding agents available from multiple teams choose service from the team with the lowest z-score for that type of call in the *Zcalls* table (bottom half of Table 3).

Overall results for the Z-score experiments are also substantial. The FCR rate increases 0.17%, AHT drops 4.52% from 194 to 185 seconds, and ASA falls 44.6% from 183 to 102 seconds. The ASA of only one call type (Inquiry) suffers (rising 30.6% from 190 to 248 seconds) while the other five call types all see their ASA reduced by at least 52%. Potential savings from these performance gains for a call center and its customers would be $19.33/day from FCR rate improvements, 4.59 hours of agent time/day from AHT decreases, and 42.92 hours of customer time/day from ASA improvements. The 4.59 hours of agent time saved equates to $137.70/day, for a total direct cost savings of $157.03/day versus $137.58/day using the AAHT strategy. A summary of these and five other strategies tested is given below in Table 4, with further explanations provided in (Stanley 2007).

Strategy	FCR Rate Cost (1)	AHT Cost (2)	ASA Cost (Hrs) (3)	Direct Costs (1+2)
AAHT	$ 23.88	$ 113.70	48.18	$ 137.58
Z-score	$ 19.33	$ 137.70	42.92	$ 157.03
FCR Rate	$ 4.55	$ 33.00	20.98	$ 37.55
AAHT & MaxTime	$ 7.98	$ 102.60	47.18	$ 110.58
AAHT + Extra Agent	$ 5.70	$ 100.50	74.77	$ 106.20
FCR Rate & MaxTime	$ 19.33	$ 34.80	21.76	$ 54.13
Modified Max Pressure	$ 15.90	$ 110.10	23.73	$ 126.00

Table 4. Summary of business savings (avoided costs) by PBR strategy.

V. CONCLUSION AND FUTURE WORK

This article has described the use of computer simulation modeling to analyze performance-based routing strategies. Results based on data from a financial call center demonstrate the potential for significant improvements in call center performance, especially ASA, by using rules based on historic performance data such as AHT and FCR rates. In addition to the rules presented here, five other PBR rules were studied and many of these showed promise and benefits relative to the base case FIFO approach.

Two directions in which this research could be extended are by (1) incorporating call abandonment, and (2) running full scale tests with real data. Modeling caller abandonment would lead to more complete conclusions about cost savings from the different strategies, and would help determine the appropriate strategy to use in a particular call center environment. However, it would require estimating the cost of an abandoned call, not necessarily an easy task. Full scale tests based entirely on data from an actual call center would also be useful to ensure that the predicted performance results would really be applicable to the system from which the data were drawn.

VI. REFERENCES

Arena, Version 10, Academic Edition (Microsoft Windows compatible CD), Rockwell Automation, Inc., Sewickley, PA, 2005.

Armistead, C., Kiely, J., Hole, L., and J. Prescott, "An Exploration of Managerial Issues in Call Centers," Managing Service Quality, 12(4), 2002, 246-256.

Dean, A. M., "The Impact of the Customer Orientation of Call Center Employees on Customers' Affective Commitment and Loyalty," Journal of Service Research, 10(2), 2007, 161-173.

Dean, A. M., "Rethinking Customer Expectations of Service Quality: Are Call Centers Different?," Journal of Service Marketing, 18(1), 2004, 60-77.

Gans, N., G. Koole, and A. Mandelbaum. 2003. "Telephone Call Centers: Tutorial, Review and Research Prospects," Manufacturing and Service Operations Management, 5(2), 79-141.

Houlihan, M., "Tensions and Variations in Call Centre Management Strategies," Human Resource Management Journal, 12(4), 2002, 67-85.

"How DHL measures its customer service performance." Customer Service Newsletter, Alexander Communications Group, 1 Dec 2005. Accessed 25 July 2007. <http://www.customerservicegroup.com/pdf/csn1205docs.pdf>.

IEX. “Skills Based Routing: An Industry Survey.” Customer Service Newsletter, 1 Jan 2003. Accessed 4 June 2007. <http://www.customerservicegroup.com/pdf/skillsbasedrouting.pdf>.

Kelton, W. David, Randall P. Sadowski, and David T. Sturrock. Simulation with Arena, 4th Edition. McGraw-Hill, New York, 2007.

Levin, Greg. “Measuring The Things That Matter -- A deep dive into seven key metrics that are most critical in gauging and securing customer satisfaction, loyalty and contact center effectiveness,” Call Center Magazine, 1 March 2007, 24-33.

Richard, Darlene D. The Customer Response Management Handbook. McGraw-Hill Australia Pty Ltd, Sydney, 2002.

Sisselman, M. E., “Value-Based Routing and Preference-Based Routing in Customer Contact Centers,” Production and Operations Management, 16(3), 2007, 277-291.

Stanley, Julie D. “Evaluating Call Routing Strategies For Heterogeneous, Traffic, Servers, and Service Failure Rates,” MBA Thesis, San Francisco State University, San Francisco, CA, August 2007.

Wallace, Rodney B. and Robert M. Saltzman. “Comparing Skill-Based Routing Call Center Simulations Using C Programming and Arena Models,” Proceedings of the 2005 Winter Simulation Conference, M. Kuhl, N. Steiger, F. Armstrong, and J. Joines, eds., December 2005, 2636-2644.