Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds .

PARTIAL CROSS TRAINING IN CALL CENTERS WITH UNCERTAIN ARRIVALS AND GLOBAL SERVICE LEVEL AGREEMENTS

Thomas R. Robbins
Terry P. Harrison

Supply Chain and Information Systems
Penn State University
University Park, PA 16802, U.S.A.

D. J. Medeiros

Industrial & Manufacturing Engineering
Penn State University
University Park, PA 16802, U.S.A.

ABSTRACT

Inbound call center operations are challenging to manage; there is considerable uncertainty in estimates of arrival rates, and the operation is often subject to strict service level constraints. This paper is motivated by work with a provider of outsourced technical support services in which most *projects* (client specific support operations) include an inbound tier one help desk subject to a monthly service level agreement (SLA). Support services are highly specialized and a significant training investment is required, an investment that is not transferable to other projects. We investigate the option of cross training a subset of agents so that they may serve calls from two separate projects, a process we refer to as partial pooling. Our paper seeks to quantity the benefits of partial pooling and characterize the conditions under which pooling is most beneficial. We find that low levels of cross training yield significant benefit.

1 INTRODUCTION

Call centers are a large and growing component of the U.S. and world economy (Gans et al. 2003). In 1999 an estimated 1.5 million workers were employed in call centers in the US alone. Large scale call centers are technically and managerially sophisticated operations and have been the subject of substantial academic research. A call center is a facility designed to support the delivery of some interactive service via telephone communications, typically an office space with multiple workstations manned by *agents* who place and receive calls (Gans et al. 2003). Call center applications include tele-marketing, customer service, help desk support, and emergency dispatch.

Our research is motivated in part by recent work with a medium sized provider of call center based technical support. While the scope of services varies from account to account, many projects are 24 x 7 support and virtually all are subject to some form of Service Level Agreement (SLA). There are multiple types of SLA, but the most common specifies a minimum level of the Telephone Service Factor (TSF). A TSF SLA specifies the proportion of calls that must be answered within a specified time. For example, an 80/120 SLA specifies that 80% of calls must be answered within 120 seconds. A very important point is that the service level applies to an extended period, typically a month, which appears to be common practice in this industry. The SLA does not define requirements for a day or an hour. So the desk is typically staffed so that at some times the service level is underachieved, sometimes overachieved, and is on target for the entire month. The outsourcing contract often specifies substantial financial penalties for failing to meet the SLA.

The key challenge involved with staffing this call center is a fixed SLA with a variable and uncertain arrival rate pattern. The number of calls presented in any ½ hour period is highly variable with multiple sources of uncertainty. In addition to day of week seasonality these call center projects also experience very significant time of day seasonality. Volume ramps up sharply in the morning with a major surge of calls between 7 and 11 AM. Volume tends to dip down around the lunch break, but a second peak occurs in the afternoon; the afternoon peak is typically lower volume then the morning peak.

Agents in this environment require extensive training on the systems they support. An agent may require up to three months of training prior to taking calls, though a more typical period is three to four weeks. The majority of this training is project specific. Once deployed it takes agents several month to reach full productivity through on the job learning and supplemental training. Because of the high costs of training, it is standard practice to train an agent to handle calls from a single project.

The staffing challenge in this call center is to find a minimal cost staffing plan that achieves the global SLA target with a high probability. The schedule must obviously be locked in before arrival rate uncertainty is revealed. While management has some recourse to adjust

manpower during the course of the day (overtime, early dismissal) these actions are generally very limited. In this paper we examine one particular hedging strategy; an approach we call *partial pooling.* In partial pooling a small subset of *super agents* are cross trained to take calls from two projects. The call center can then be viewed as a skills-based routing (SBR) model with two skills. Super agents possess both skills, while *base agents* have only one skill set. It is clear that cross training all agents will increase the service level of the call center for a fixed level of staffing. Our hypothesis is that cross training a small number of agents can deliver a substantial portion of the benefit and our objective is to find the level of cross training that minimizes staffing costs, while satisfying the service level constraint with high probability.

2 LITERATURE REVIEW

A summary of the call center oriented literature is provided in (Gans et al. 2003). This detailed tutorial and review provides a thorough overview of call center operations, terminology and research. The paper summarizes the key academic research on multiple areas related to call center research, including capacity management and scheduling. Detailed analysis of models with abandonment, the Erlang A models, are provided in (Garnett et al. 2002; Mandelbaum and Zeltyn 2004; Whitt 2006). Empirical analysis of call center data is provided in a series of related papers (Mandelbaum et al. 2001; Brown et al. 2005). These papers contain a detailed statistical analysis of data from a small call center. The authors test many common assumptions used in queuing models, and find among other things that talk time in a call center follows a lognormal distribution, rather than the commonly assumed exponential distribution. Some call centers exhibit arrival behavior that has higher variability than a Poisson process and large correlations between periods in a day (Avramidis et al., 2004).

A paper very similar in concept to ours is (Wallace and Whitt 2005). (We refer to this paper as W&W) In the W&W model there are 6 call types and every agent is trained to handle a fixed number of those types. The authors use a simulation based optimization model to find the ideal cross training level. The paper's key insight is that a low level of cross training provides "most" of the benefit. Specifically, they find that training every agent in two skills provides the bulk of the benefit, while additional training has a relatively low payoff. Although the general finding in our paper is similar, e.g. small levels of cross training give the majority of the benefit, the models are very different. While their best solution has every agent cross trained in two skills, our model assumes that only a small proportion of agents are cross trained. In our scenario cross training is very expensive and 100% cross training is not practical. The W&W model also ignores abandonment, an important consideration in our situation.

Other approaches to staffing multiskill call centers include stochastic fluid models (Harrison and Zeevi 2005), search methods with a loss-delay approximation (Avramidis et al. 2006) and a cutting-plane algorithm combined with simulation (Cezik and L'Ecuyer 2006). These approaches are designed for large call centers with many customer classes and skill classifications. They minimize the cost of staffing the center subject to some service level constraints, where the cost of an agent is a function of the number of skills that agent possesses. They solve the problem for a single time period assuming steady state demand; the simulation approach also includes abandonment and fixed routing based on customer and staff priorities.

3 POOLING MODEL

In this section we introduce our model of partial pooling. We assume that in the baseline case the call center is segregated by project and each project acts a separate Erlang-A queuing system. Each project *i* receives calls that arrive with rate λ_i. Associated with each call is a average talk time denoted as $1/\mu_i$. We also assume that callers have exponentially distributed patience with mean $1/\theta_i$. The patience parameter represents the time a caller is willing to wait on hold. Each caller will abandon the queue (hang up) if not server by their patience parameter. For our purposes the key performance metric for this queue is the Telephone Service Factor is defined as

$$TSF = P\{W \le T\}$$

The steady state staffing decision then involves forecasting the arrival rate λ_i and setting the staff level to achieve the specified SLA with an acceptable level of probability.

Our analysis is based on a modification to existing practice; specifically cross training a number of agents to support two projects. In our model we assume that the skills based routing system is configured as follows

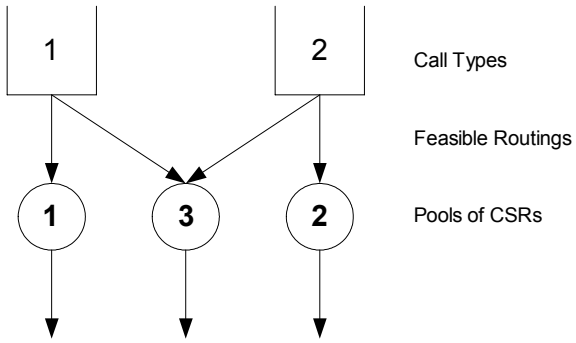

Figure 1: Basic routing structure

We have two call types, one for each project, and three agent pools. Pool 1 has skill 1 and can service calls of type

1. Similarly pool 2 services calls of type 2. Pool 3 is cross trained and can service calls of either type. We implement a very simple routing model. An incoming call is routed to a base agent if one is available. Only in the case where all base agents are busy is the call routed to a super agent. When base agents become available they take the longest waiting caller from their respective queue. If no calls are waiting they become idle. When a super agent becomes available they take the call from the largest queue.

4 PARTIAL POOLING IN STEADY STATE

4.1 The TSF Response Function

In the single queue, single resource pool case, we have an analytical expression for the service level as a function of arrival rates and staffing and we can easily generate a plot of the TSF as a function of staffing. The TSF function generates an s-shaped curve, with the service level improving rapidly at first, then leveling off and finally showing declining returns to incremental staffing. In the pooling case the situation is considerably more complicated. There are no known analytical expressions available to calculate the service level. Based on intuition we expect the service level is increasing in the number of base agents and the number of super agents. To verify this intuition we use simulation to create the following graphical representation of the TSF as a function of the number of agents.

In this simulation we assume that each queue receives calls at a rate of 100 calls per hour, that in each case talk time averages 12 minutes, callers have an average patience of 350 seconds, and the service level is based on a 120 second hold time. We vary the number of agents assigned to each base pool and the number of agents assigned to the super agent pool independently. For each staffing combination we simulate operations for two days, and perform 25 replications.

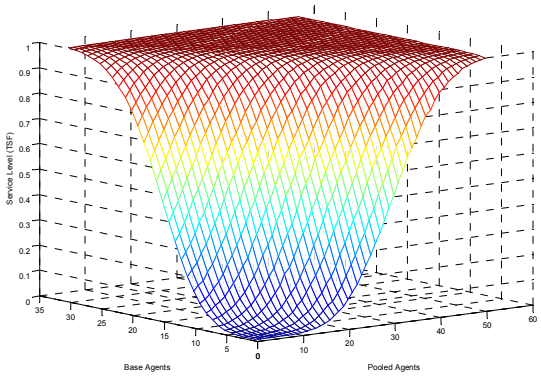

Figure 2: The TSF response function for a partial pooling model

The graph illustrates a large plateau of 100% TSF when the total number of agents is large. Similarly a small plateau at 0% TSF exists when the total number of agents is small. In between the surface exhibits an S shaped profile. Figure 3 is a contour plot of this data in two dimensions. The contour plot shows a series of iso-service level lines, agent combinations that deliver the same service level.

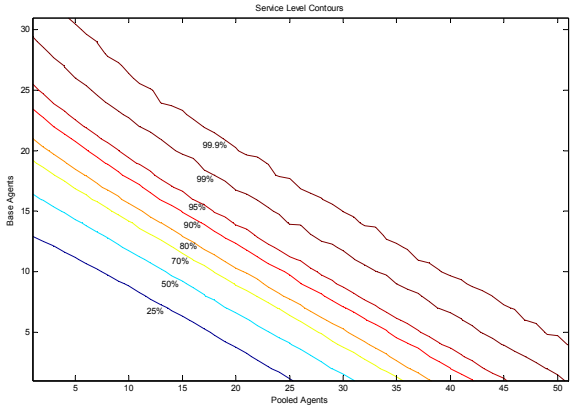

Figure 3: The pooled TSF contour plot

So for example, to achieve a 95% service level we need roughly 25 agents in each pool or 50 agents overall. However, in a pure pooled mode the same service level can be achieved with a total of only 45 pooled agents.

Though difficult to see, close inspection reveals that the iso-service lines are not straight, but have a convex *bowed* shape. This is further illustrated in the next figure where we show the 80% TSF contour with a line connecting the end points. The convexity of the contour implies that the cost minimizing combination of pooled and base agents may be in the interior.

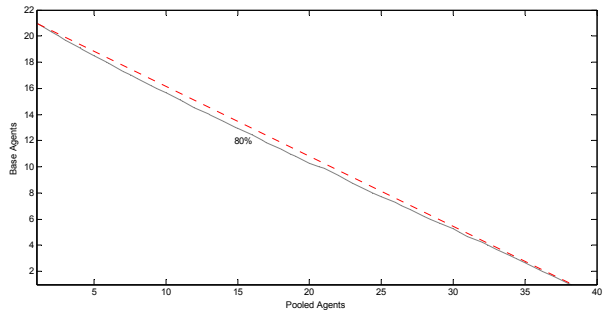

Figure 4: 80% TSF contour

4.2 Symmetric Projects

In this experiment we test the impact of pooling on steady state performance with symmetric projects. Consider two statistically identical projects each staffed with 36 agents

and receiving calls at a constant rate λ. Talk time has an exponential distribution with mean 12 minutes and the mean time to abandon is 350 seconds. The service level is measured against a two minute hold time. We evaluate the situation where the total number of agents remains constant, but each project contributes between 0 and 36 agents to the pool. The first graph shows the service level for each level of pooling when λ is 200 calls per hour. In the second graph we plot the abandonment rate. In each case we plot TSF and abandonment rate for one of the projects. (Because of the symmetric nature of the model, each project has the same curve.) The data was generated by simulating five days of operations over 50 replications. In each curve we show the sample average along with a 90% confidence interval.

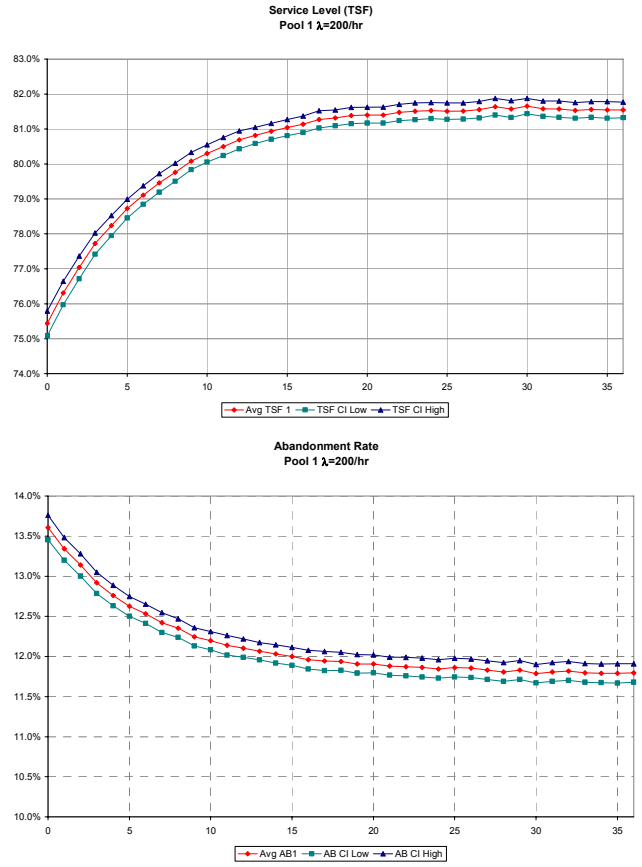

Figure 5: Impact of pooling on TSF and abandonment with fixed staffing levels

These graphs reveal that a small level of pooling yields improvement, but that the return on cross training declines rapidly. In each case cross training 10 agents provides the bulk of the benefit and cross training beyond 15 agents provides very limited benefits. We have repeated this test with arrival rates of 180 and 220 calls per hour and obtained similar results. In each case cross training can boost TSF by 5%-6%, while the biggest improvement is in the medium volume (200/hr) case. Abandonment is decreased by about 1% in the high volume case, 1.8% in the middle case, and 2.3% in the slow case.

4.3 Steady State Differential Rates

The previous analysis reveals that moderate benefits are achieved when agents are cross trained, and the amount of improvement depends on the spare capacity in the system. However in that analysis both projects had the same arrival rate. A more interesting case occurs when the arrival rates are different as may be the case if rates are subject to forecast error. In the next analysis we allow arrival rates to vary independently from target by ± 10%. Total staffing is fixed at 72, so that in the no pooling case each project has 36 agents, a staff level that results in an approximately 76% service level with no pooling. The following tables summarize the resulting TSF measures under various arrival rate combinations. The numbers along the top (0-35) indicate the number of agents pooled.

In Table 1 we examine the impact on the combined TSF, and the TSF of each individual project. We see that the overall TSF is always improved by pooling, and the degree of improvement is based on the amount of spare capacity in the system. When both projects have below plan volume, the overall TSF is improved by 2.8% with just five agents. If both projects have above plan volume, the TSF also improves by 2.8%. The biggest gain comes when the projects have differential rates; when one project is low and the other high we get a 5.2% gain in overall TSF. The improvement quickly drops off with the number of agents cross trained; the most benefit comes from the first few agents. Cross training beyond 15 agents yields results that are not meaningful, and in many cases are not statistically different from zero.

Table 1: Impact of pooling on overall TSF. * indicates statistical significance at the .9 level

TSF Total

λ1	λ2	0	5	10	15	20	25	30	35
180	180	87.9%	90.8%	91.9%	92.5%	92.7%	92.8%	92.9%	92.8%
180	200	81.5%	85.3%	86.9%	87.6%	88.0%	88.1%	88.1%	88.0%
180	220	73.6%	78.8%	80.9%	81.9%	82.4%	82.5%	82.5%	82.5%
200	200	76.1%	79.3%	80.8%	81.6%	81.9%	82.0%	82.1%	81.9%
200	220	68.8%	72.3%	74.2%	75.0%	75.3%	75.6%	75.6%	75.5%
220	220	62.3%	65.1%	66.6%	67.5%	67.8%	68.0%	68.0%	68.1%

Δ TSF Total

λ1	λ2	5	10	15	20	25	30	35
180	180	2.8% *	1.1% *	0.6% *	0.2% *	0.1% *	0.0%	0.0%
180	200	3.8% *	1.6% *	0.7% *	0.3% *	0.1% *	0.0%	0.0%
180	220	5.2% *	2.2% *	1.0% *	0.5% *	0.1% *	0.0%	-0.1% *
200	200	3.2% *	1.5% *	0.8% *	0.3% *	0.1% *	0.1% *	-0.2% *
200	220	3.6% *	1.9% *	0.8% *	0.3% *	0.2% *	0.0%	-0.1% *
220	220	2.8% *	1.5% *	0.9% *	0.3% *	0.3% *	0.0%	0.1%

The results are even more interesting when we examine the data at the individual project level (Tables 2-3). When each project has a similar arrival rate the benefits are distributed evenly. But it is when the arrival rates are different that the maximum gain occurs; and that gain accrues

disproportionately to the understaffed project. When volumes are at opposite extremes, the understaffed project receives a benefit of an 11% boost in TSF from only 5 cross trained agents. Cross training of 10 agents increases TSF by another 10 points raising TSF to nearly 80%. In the case of significant mismatch the overstaffed project may suffer degradation in performance, but this decline is significantly smaller then the boost to the other project and aggregate TSF always increases. The most significant case is when volumes have a maximum mismatch and the overstaffed project's TSF declines by 2.2% with 5 agents cross trained. Note however that this project had a baseline TSF of 86%, well over the standard target of 80%. This result does however raise a caution for pooling projects with very high (90%) TSF targets. In the case of a smaller mismatch the degradation was very moderate, about 0.9% with 10 agents cross trained, where the busy project may see an improvement on the order of four points from only 5 cross trained agents. We obtained similar results for the abandonment rate for each project; pooling reduces the maximum wait time callers face, and therefore reduces the proportion of callers kept on hold past their patience level. The improvement is the most significant when a capacity mismatch occurs.

Table 2: Impact of Pooling on low volume project TSF

		TSF Pool 1							
λ1	λ2	0	5	10	15	20	25	30	35
180	180	86.8%	90.4%	91.7%	92.4%	92.7%	92.8%	92.8%	92.8%
180	200	86.8%	87.5%	87.7%	87.8%	87.7%	87.6%	87.4%	87.3%
180	220	86.8%	84.7%	82.9%	82.1%	81.4%	80.9%	80.6%	80.2%
200	200	75.5%	78.9%	80.5%	81.4%	81.8%	82.0%	82.0%	82.0%
200	220	75.5%	75.3%	75.1%	74.9%	74.7%	74.4%	74.3%	74.0%
220	220	61.9%	65.0%	66.4%	67.6%	68.0%	68.1%	68.2%	68.2%

		Δ TSF Pool 1						
λ1	λ2	5	10	15	20	25	30	35
180	180	3.5% *	1.3% *	0.7% *	0.2% *	0.1% *	0.0%	0.0%
180	200	0.7% *	0.2% *	0.1%	0.0%	-0.1% *	-0.2% *	-0.1% *
180	220	-2.2% *	-1.7% *	-0.8% *	-0.7% *	-0.5% *	-0.3% *	-0.4% *
200	200	3.4% *	1.6% *	1.0% *	0.4% *	0.1% *	0.1%	-0.1% *
200	220	-0.2%	-0.2%	-0.3% *	-0.2% *	-0.2% *	-0.1% *	-0.2% *
220	220	3.1% *	1.5% *	1.1% *	0.4% *	0.2% *	0.1%	0.0%

Table 3: Impact of Pooling on high volume project TSF

		TSF Pool 2							
λ1	λ2	0	5	10	15	20	25	30	35
180	180	89.0%	91.1%	92.1%	92.5%	92.8%	92.9%	92.9%	92.8%
180	200	76.7%	83.3%	86.1%	87.5%	88.2%	88.5%	88.7%	88.7%
180	220	62.7%	73.9%	79.3%	81.8%	83.2%	83.8%	84.1%	84.3%
200	200	76.7%	79.7%	81.0%	81.7%	82.0%	82.0%	82.2%	81.9%
200	220	62.7%	69.6%	73.4%	75.1%	75.9%	76.6%	76.8%	76.8%
220	220	62.7%	65.2%	66.7%	67.4%	67.6%	67.9%	67.8%	68.0%

		Δ TSF Pool 2						
λ1	λ2	5	10	15	20	25	30	35
180	180	2.1% *	0.9% *	0.4% *	0.3% *	0.1% *	0.0%	-0.1% *
180	200	6.6% *	2.8% *	1.4% *	0.7% *	0.4% *	0.1% *	0.0%
180	220	11.3% *	5.3% *	2.5% *	1.5% *	0.6% *	0.3% *	0.2% *
200	200	3.0% *	1.3% *	0.6% *	0.3% *	0.0%	0.1% *	-0.2% *
200	220	7.0% *	3.8% *	1.7% *	0.9% *	0.7% *	0.2% *	0.0%
220	220	2.6% *	1.5% *	0.6% *	0.2% *	0.3% *	-0.1%	0.2% *

Overall this analysis shows that partial pooling yields substantial benefits in steady state. The improvement is the greatest when a capacity mismatch occurs and the under capacity project receives the greater benefit. In the next section we examine how arrival rate uncertainty impacts the pooling analysis.

4.4 Steady State but Uncertain Arrival Rate

In this analysis we continue to examine the impact of pooling when projects have a constant rate, but we now allow for uncertainty in the arrival rate. Specifically we assume that the calls in each pool will arrive with a constant rate, but the realized rate is a random variable. Assume that the arrival rates are independent and identically distributed normal random variables with mean 200 and standard deviation 20. We examine how partial pooling impacts the expected TSF and abandonment rate. The following graph present the results of a simulation experiment.

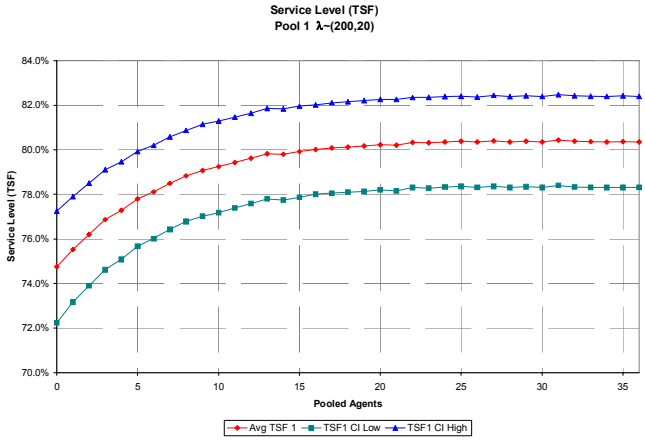

Figure 6: Impact of pooling on TSF with fixed staffing levels and uncertain arrivals- TSF confidence level

The service curve generated in this case is almost identical to the plots for steady state arrivals at 200 calls/hr but the TSF level is lower in the uncertain case: 74.2% vs. 75.4% with no pooling and 80.5% vs. 81.5% in the full pooling case. Although not a major shift, this illustrates one of the effects of arrival rate uncertainty. Because of the nature of the TSF curve the effect of volume changes is not proportional; higher volume causes a larger shift in the resulting service level then lower volume. So even if volume varies around the mean symmetrically, the resulting TSF will be lower in the uncertain case than the corresponding mean value case. Similar results were found for abandonment: the curves for the uncertain case have a similar shape as the certain case, with a moderately higher abandonment rate at all levels.

An interesting phenomenon is illustrated in Figures 7 and 8. In Figure 7 we see that the standard deviation of the

overall (combined) service level is essentially unaffected by pooling, remaining at a roughly constant level just over 8%.

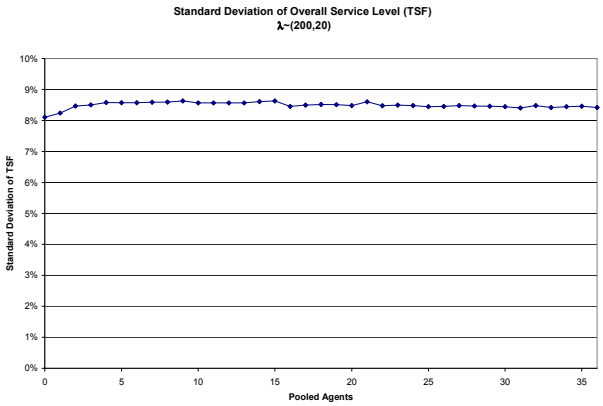

Figure 7: Standard deviation of the overall TSF

Figure 8 however reveals the standard deviation of the service level for pool one decreases as the pooling level increases, at least for the first few pooled agents.

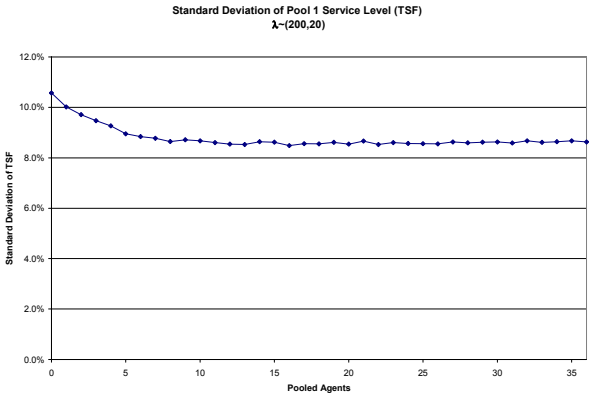

Figure 8: Standard deviation of the individual project TSF

In the case of no pooling the service level in each pool is independent from the service level in the other pool. As pooling increases the service levels in each pool become dependent random variables.

5 SUMMARY AND CONCLUSIONS

This paper examines the issue of partial cross training in call centers that must achieve a global service level. We deal with the case where each project has a single skill-set; this may not be ideal but is the case in several projects we analyzed. The current analysis focuses on pooling in steady state operations and characterizes the improvement from partial pooling. Our ongoing research is focused on finding the optimal cross training level; that is the level at which the benefit of cross training is balanced by the extra cost of cross training. Our ongoing research is focused on developing algorithms to find the optimal cross training level with both steady state and non stationary arrival rates.

REFERENCES

Avramadis, A., A. Deslauriers, and P. L'Ecuyer 2004. Modeling daily arrivals to a telephone call center. *Management Science* 50(7):896–908.

Avramidis, A., W. Chan, and P. L'Ecuyer 2007. Staffing multi-skill call centers via search methods and a performance approximation. Working Paper [online] 45p. Available via <www.iro.umontreal.ca/~lecuyer/papers.html> [accessed July 5, 2007].

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, S. Haipeng, S. Zeltyn and L. Zhao 2005. Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association* 100(469):36-50.

Cezik, M, and P. L'Ecuyer 2007. Staffing multiskill call centers via linear programming and simulation. Working Paper [online] 34p. Available via <www.iro.umontreal.ca/~lecuyer/papers.html> [accessed July 5, 2007].

Gans, N., G. Koole and A. Mandelbaum 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79-141.

Garnett, O., A. Mandelbaum and M. I. Reiman 2002. Designing a Call Center with impatient customers. *Manufacturing & Service Operations Management* 4(3):208-227.

Harrison, J. and A. Zeevi 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management* 7(1):20-36.

Mandelbaum, A. and S. Zeltyn. 2004. Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers. Working Paper [online] 29p. Available via <iew3.technion.ac.il/serveng/References/references.html> [accessed March 10, 2006].

Wallace, R. B. and W. Whitt 2005. A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management* 7(4):276-294.

Whitt, W. 2006. Sensitivity of Performance in the Erlang A Model to Changes in the Model Parameters. *Operations Research* 54(2):247-260.

AUTHOR BIOGRAPHIES

THOMAS ROBBINS recently earned a PhD in Business Administration and Operations Research from Penn State University. He worked in professional services for approximately 18 years before entering the Ph.D. program. He served as a partner at Ernst and Young, and a Vice President at CGEY and Aztec Software. He holds an MBA from Case Western Reserve and a BSEE from Penn State. His research interests include service operations and stochastic optimization. His email address is <trr147@psu.edu>.

D. J. MEDEIROS is Associate Professor of Industrial Engineering at Penn State University. She holds a Ph.D. and M.S.I.E from Purdue University and a B.S.I.E. from the University of Massachusetts at Amherst. She has served as track coordinator, Proceedings Editor, and Program Chair for WSC. Her research interests include manufacturing systems control and CAD/CAM. She is a member of IIE and SME. Her email address is <djm3@psu.edu>.

TERRY P. HARRISON is the Earl P. Strong Executive Education Professor of Business and Professor of Supply Chain and Information Systems at Penn State University. He holds a Ph.D. and M.S. degree in Management Science from the University of Tennessee and a B.S in Forest Science from Penn State. He was formally the Editor in Chief of Interfaces and is currently Vice President of Publications for Informs. His mail address is <tharrison@psu.edu>.