Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

EVALUATING THE FIT OF THE ERLANG A MODEL IN HIGH TRAFFIC CALL CENTERS

Thomas R. Robbins

College of Business
East Carolina University
Greenville, NC 28590, USA

ABSTRACT

We consider the Erlang A model, a queuing model often applied to analyze call center performance. While not a new model, Erlang A is becoming a popular alternative to the widely used Erlang C model. In this paper we analyze the accuracy of Erlang A predictions in high traffic environments, a situation where the Erlang C model is not applicable. Our findings indicate that in this high traffic region the Erlang A model is subject to a moderate to high level of error that has a strong pessimistic bias; that is the system tends to perform better than predicted. This is in sharp contrast to lower volume scenarios where the model tends to be optimistically biased. We find that in addition to utilization, the model is most sensitive to arrival rate uncertainty and balking.

1 INTRODUCTION

Call centers are examples of queuing systems; calls arrive, wait in a virtual queue, and are then serviced by an agent. Call centers are often modeled as M/M/N queuing systems, or in industry standard terminology - the Erlang C model. The Erlang C model makes many assumptions that are questionable in the context of a call center environment. Most significantly, Erlang C assumes that all callers wait as long as necessary for service without abandoning. An increasingly popular alternative model is the Erlang A model, an extension of the Erlang C model that allows for caller abandonment. While many papers have noted the deficiencies of the Erlang C model and advocated the use of the Erlang A model, a systematic analysis of the error associated with each model is lacking. Our paper seeks to close this gap in the literature.

In this paper we analyze call center performance in a region the Erlang C model is not applicable, the high traffic environment often referred to the efficiency-driven regime. In this regime the call center lacks the capacity to handle all calls presented. Faced with long waiting times, a significant proportion of callers abandon the queue. This abandonment brings the capacity required in line with the capacity available. The purpose of this study is to evaluate the performance of the Erlang A models in this environment, and to determine if it differs substantially from Erlang A performance in lower utilization environments. We conduct this analysis by performing a detailed simulation study. We develop a simulation model to predict steady state expected system performance based on a realistic set of modeling assumptions as identified in the literature. Our findings indicate that the Erlang A model may be subject to reasonably high levels of error in this region and further that the model's error has significantly different characteristics in the high traffic region.

The remainder of this paper is organized as follows. In Section 2 we review the Erlang C and Erlang A models and highlight the relevant literature. In Section 3 we present a general model of a steady state call center environment and review the simulation model we developed to evaluate it. In Section 4 we evaluate the performance of the Erlang A model. We conclude in Section 5 with summary observations and identify future research questions.

2 QUEUING MODELS AND THE ASSOCIATED LITERATURE

2.1 The Baseline Model – Erlang C

The literature focused on call centers is quite large, with thorough and comprehensive reviews provided in (Gans, Koole, and Mandelbaum 2003) and (Aksin, Armony, and Mehrotra 2007). Empirical analysis of call center data is given in (Brown et al. 2005).

The most common queuing model used for inbound call centers is the Erlang C model (Brown et al. 2005; Gans, Koole, and Mandelbaum 2003). The Erlang C model (M/M/N queue) is a very simple multi-server queuing system. Calls arrive according to a Poisson process at an average rate of λ. All calls that enter the queue are serviced by a pool of n homogeneous (statistically identical) agents at an average rate of n. Service times follow an exponential distribution with a mean service time of μ^{-1}. The *offered utilization* is defined as $\rho \triangleq \lambda/(n\mu)$. Given the assumption that all calls are serviced, the traffic intensity must be strictly less than one or the queue grows without bound. The proportion of callers that must queue prior to service, ***ProbWait***, is a basic measure of system performance. Another relevant performance measure for call centers managers is the ***Average Speed to Answer*** (ASA). A third important performance metric for call center managers is the ***Telephone Service Factor*** (TSF), also called the "service level." The TSF is the fraction of calls presented which are eventually serviced and for which the delay is below a specified level. For example, a call center may report the TSF as the percent of callers on hold less than 30 seconds. A fourth performance metric monitored by call center managers is the ***Abandonment Rate*****;** the proportion of all calls that leave the queue (hang up) prior to service. Abandonment rates cannot be estimated directly using the Erlang C model because the model assumes no abandonment occurs.

The Erlang C model assumes also that calls arrive according to a Poisson process. The interarrival time is a random variable drawn from an exponential distribution with a known arrival rate. Several authors assert that the assumption of a known arrival rate is problematic. Both major call center reviews (Gans, Koole *et al.* 2003; Aksin, Armony *et al.* 2007) have sections devoted to arrival rate uncertainty. Brown et al. (2005) perform a detailed empirical analysis of call center data and suggest that the arrival rate should be modeled as a stochastic process. Several other authors also argue that call center arrivals follow a doubly stochastic process; a Poisson process where the arrival rate is itself a random variable (Chen and Henderson 2001; Whitt 2006c; Aksin, Armony, and Mehrotra 2007). Arrival rate uncertainty may exist for multiple reasons that are not captured in forecasts. Robbins (2007) compares four months of week-day forecasts to actual call volume for 11 call center projects. He finds that the average forecast error exceeds 10% for 8 of 11 projects, and 25% for 4 of 11 projects. The standard deviation of the daily forecast to actual ratio exceeds 10% for all 11 projects. Steckley, Henderson, and Mehrotra (2009) compare forecasted and actual volumes for nine weeks of data taken from four call centers. They show that the forecasting errors are large and modeling arrivals as a Poisson process with the forecasted call volume as the arrival rate can introduce significant error. Robbins, Medeiros, and Dum (2006) use simulation analysis to evaluate the impact of forecast error on performance measures demonstrating the significant impact forecast error can have on system performance. The Erlang C model also assumes that the service time follows an exponential distribution. However, empirical analysis suggests that the exponential distribution is a relatively poor fit for service times. Most detailed analysis of service time distributions find that the lognormal distribution is a better fit (Brown et al. 2005; Gans, Koole, and Mandelbaum 2003; Mandelbaum A., Sakov A. , and S. 2001). Finally, the Erlang C model assumes that agents are *homogeneous*. Empirical evidence supports the notion that agents are heterogeneous with different service time distributions (Armony and Ward 2008) (Robbins 2007).

2.2 The Erlang A Extension

Given the prevalence of caller abandonment in modern call centers, the *no abandonment* assumption of the Erlang C model may be problematic. Unfortunately, models that allow for abandonment are significantly

more complex and difficult to characterize. The simplest abandonment model is the $M/M/N+M$, or Erlang A model. The model was originally presented by Palm in a 1946 paper written in Swedish. It was presented in English in Palm (1957). The Erlang A model is presented in detail in Gans, Koole, and Mandelbaum (2003) and Mandelbaum and Zeltyn (2007).

In the Erlang A model each caller posses an exponentially distributed *patience time* with mean θ^{-1}. If the offered waiting time, the time a caller with infinite patience would be required to wait, exceeds the customer's patience time, the caller will abandon the queue and hang up (Mandelbaum and Zeltyn 2007). While the exponentially distributed patience time makes the calculations tractable, they are by no means straightforward. Details on how to calculate performance metrics for the Erlang A model are provided in Mandelbaum and Zeltyn (2009). Garnett, Mandelbaum, and Reiman (2002) outlines a method for an exact calculation of the Erlang A performance metrics, and also provides approximations based on an asymptotic analysis of the queue. Whitt (2006a) develops deterministic fluid models to provide simple first-order performance descriptions for multiserver queues with abandonment under heavy loads.

The inclusion of abandonment has a profound effect on the performance of the queuing system, the specifics of which are discussed in detail in Garnett, Mandelbaum, and Reiman (2002). First of all, the issue of system stability is no longer a concern. Furthermore, even very low levels of caller abandonment can dramatically alter system performance. Comparisons of Erlang C and Erlang A models are developed in Mandelbaum and Zeltyn (2007) and Garnett, Mandelbaum, and Reiman (2002). Whitt (2005) examines the fit of the Erlang A model. Whitt (2006b) examines the sensitivity of the Erlang A model to changes in the model parameters. Several papers examine staffing and scheduling issues in call centers where abandonment is allowed (Avramidis et al. 2007; Bassamboo, Harrison, and Zeevi 2005; Robbins and Harrison 2010). In order to develop a tractable model, the Erlang A model assumes an exponentially distributed patience. Brown et al. (2005) examine abandonment and a customer's willingness to wait in detail and find significant deviations from this assumption. Several other studies of patience curves have concluded that patience can be best modeled as a Weibull distribution (Gans, Koole, and Mandelbaum 2003).

Robbins, Medeiros, and Harrison (2010) examines the fit of the Erlang C model in a realistic call center setting. They compare the Erlang model predictions with the results of a simulation study that relaxes several of the key assumptions in the Erlang models. Their study evaluates the models over an experimental region that encompasses offered utilization rates that range from 65% to 95%, environments often referred to as the *quality-driven and quality and efficiency-driven* (QED) regimes (Gans, Koole, and Mandelbaum 2003). That study finds that while the Erlang A model is in general more accurate than Erlang C it tends to be optimistically biased when the arrival rate is uncertain. This paper builds on that analysis by examining the performance of Erlang A when the offered utilization level exceeds 100%.

3 CALL CENTER SIMULATION

3.1 The Modified Model

In this section we present a revised model of a call center, relaxing several key assumptions discussed previously. In our model calls arrive at a call center according to a Poisson process. Calls are forecasted to arrive at an average rate of $\hat{\lambda}$. The realized arrival rate is λ, where λ is a normally distributed random variable with mean $\hat{\lambda}$ and standard deviation σ_λ. The time required to process a call by an average agent is a lognormally distributed random variable with mean μ^{-1} and standard deviation σ_μ. Arriving calls are routed to the agent who has been idle for the longest time if one is available. If all agents are busy the call is place in a FCFS queue. When placed in queue a proportion of callers will balk; *i.e.* immediately hang up. Callers who join the queue have a patience time that follows a Weibull distribution with parameters α and β. If wait time exceeds their patience time the caller will abandon. Calls are serviced by agents who have variable relative productivity r_i. An agent with a relative productivity level of 1 serves calls at the

average rate. An agent with a relative productivity level of 1.5 serves calls at 1.5 times the average rate. Agent productivity is assumed to be a normally distributed random variable with a mean of 1 and a standard deviation of σ_r.

3.2 Experimental Design

In order to evaluate the performance of the Erlang A models against the simulation model, we conduct a series of designed experiments. Based on the assumptions for our call center discussed previously, we define the following set of nine experimental factors.

Table 1: Experimental Factors in the model.

	Factor	**Low**	**High**
1	Number of Agents	10	100
2	Offered Utilization ($\hat{\rho}$)	100%	200%
3	Talk Time (mins)	2	20
4	Patience β	60	600
5	Forecast Error CV (ARCV)	0	.2
6	Patience α	.75	1.25
7	Talk time CV	.75	1.25
8	Probability of Balking	0	.25
9	Agent Productivity Standard Deviation	0	.15

The forecasted arrival rate in the simulation is a quantity derived from other experimental factors by

$$\hat{\lambda} = \hat{\rho} N \mu \qquad (1)$$

Given the relatively large number of experimental factors, a well-designed experimental approach is required to efficiently evaluate the experimental region. A standard approach to designing computer simulation experiments is to employ either a full or fractional factorial design (Law 2007). However, the factorial model only evaluates corner points of the experimental region and implicitly assumes that responses are linear in the design space. We chose instead to implement a Space Filling Design based on Latin Hypercube Sampling as discussed in (Santner, Williams, and Notz 2003). Given a set of d experimental factors and a desired sample of n points, the experimental region is divided into n^d cells. A sample of n cells is selected in such a way that the centers of these cells are uniformly spread when projected onto each of the d axes of the design space. We chose our design point as the center of each selected cell. This experimental design allows us to select an arbitrary number of points for any experiment.

3.3 Simulation Model

Our call center model is evaluated using a straightforward discrete event simulation model coded in Visual Basic. The purpose of the model is to predict the long term, steady state behavior of the queuing system. The model generates random numbers using a combined multiple recursive generator (CMRG) based on the Mrg32k3a generator described in (L'Ecuyer 1999). Common random numbers are used across design points to reduce output variance. To reduce any start up bias we use a warm up period of 5,000 calls, after which all statistics are reset. The model is then run for an evaluation period of 25,000 calls and summary statistics are collected. For each design point we repeat this process for 500 replications and report the average value across replications. Our primary analysis is based on an experiment with 1,000 design points.

The specific process for each replication is as follows. The input factors are chosen based on the experimental design. The average arrival rate is calculated based on the specified talk time, number of agents, and offered utilization rate according to equation (1). A random number is drawn and the realized arrival rate is set based on the probability distribution of the forecast error. That arrival rate is then used to generate Poisson arrivals for the replication. Agent productivities are generated using a normal distribution with mean one and standard deviation σ_p. Each new call generated includes an exponentially distributed interarrival time, a lognormally distributed nominal talk time, a Weibull distributed time before abandonment, and a Bernoulli distributed balking indicator. When the call arrives it is assigned to the longest idle agent, or placed in the queue if all agents are busy. If sent to the queue the simulation model checks the balking indicator. If the call has been identified as a balker it is immediately abandoned, if not an abandonment event is scheduled based on the realized time to abandon. Once the call has been assigned to an agent, the realized talk time is calculated as the product of the nominal talk time and the agent's productivity. The agent is committed for the realized talk time. When the call completes the agent processes the next call from the queue, or if no calls are queued becomes idle. If a call is processed prior to its time to abandon, the abandonment event is cancelled. If not, the call is abandoned and removed from the queue. Over the course of the simulation we collect statistics on the proportion of customers forced to wait, the average speed to answer, the abandonment rate, and the TSF defined as the proportion of callers waiting less than 30 seconds.

After all replications of the design point have been executed the results are compared to the theoretical predictions of the Erlang A model. We wish to eliminate any approximation errors in our comparison, so rather than use an approximate calculation for the Erlang A model we rerun the simulation configured to be consistent with the Erlang A model assumptions, *i.e.* no balking, homogeneous agents, exponential talk time and exponential patience. The simulation is run using common random numbers from the original simulation. We feel that this approach allows us to focus on the error associated with the Erlang A assumptions, rather than the numerical issues associated with estimating Erlang A performance measures. The second challenge is how to set the patience parameter for the Erlang A calculation. Recall that this parameter is not directly observable since data is heavily censored. Since we are attempting to fit the Erlang A model to observed data, we approximate the Erlang A parameter θ as in (Gans, Koole, and Mandelbaum 2003) and (Brown et al. 2005) by $\theta = P\{Abandon\}/E[Wait]$.

4 EXPERIMENTAL ANALYSIS

4.1 Summary Observations

Based on the results of this analysis we identify the following summary observations.

- Errors are correlated in a statistically significant manner. ProbWait error is strongly correlated with the TSF error and utilization error; correlation with the ASA error is moderate, correlation with Abandonment error is weak.
- ProbWait errors are moderate to large. The average error is 5%, but errors are recorded as high as 24%.
- ProbWait errors are positively (pessimistically) skewed; the system tends to perform better than predicted.

4.2 Correlation and Magnitude of Errors

Table 2: A correlation matrix of measurement errors.

	Prob Wait Error	*ASA Error*	*TSF Error*	*Abandonment Rate Error*	*Utilization Error*
Prob Wait Error	1.000				
ASA Error	-.385	1.000			
TSF Error	-.645	.093	1.000		
Abandonment Rate Error	.066	.352	-.416	1.000	
Utilization Error	.690	-.421	-.192	-.282	1.000

1000 sample size

± .062 critical value .05 (two-tail)
± .081 critical value .01 (two-tail)

The correlations between the errors and performance measurements are all statistically significant. The correlation between the key performance metrics ProbWait and TSF is reasonably strong. These measures are negatively correlated; a higher proportion of callers waiting implies a lower TSF. ProbWait error correlates strongly with the error in the utilization calculation. ProbWait correlates less strongly with ASA, and very weakly with the error in the abandonment calculation. The correlation between ProbWait and ASA, two measures in which a lower value indicates better system performance, are negatively correlated.

The relationship between the ProbWait and ASA errors are further illustrated in the Figure **1**. ASA errors tend to be positive. We term this a pessimistic or conservative error since the system tends to perform better than predicted. Negative, or optimistic, errors in which the system performs worse than expected are possible but the magnitudes are relatively small. ProbWait errors also tend to be positive and these errors are also pessimistic as the system is behaving better than predicted with a smaller proportion of calls queuing. It is interesting to note that large ASA errors correspond to relatively small ProbWait errors, and large ProbWait errors correspond to relatively small SA errors.

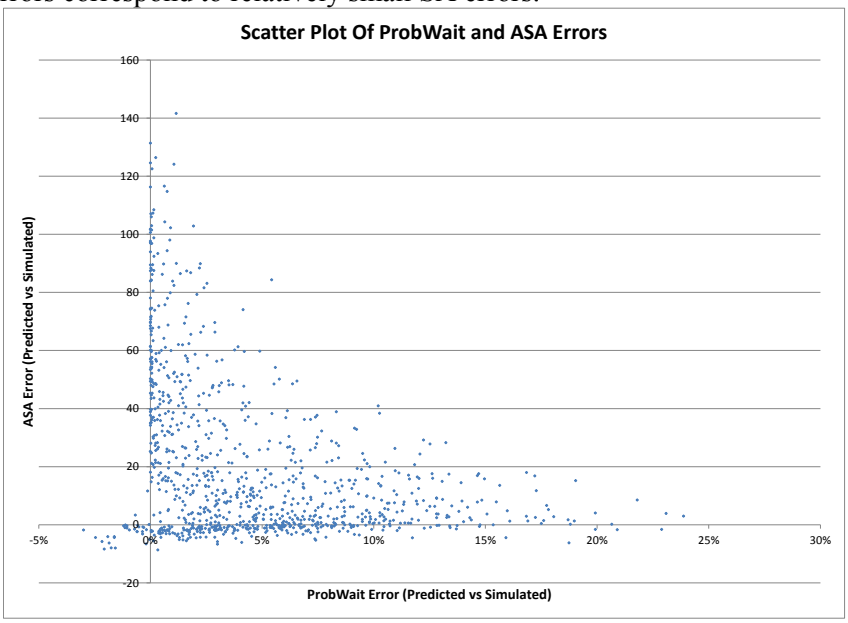

Figure 1: A scatter plot of ASA and ProbWait errors.

This effect is further illustrated in Figure **2** where we present a scatter plot of the proportional error, the difference in predicted and observed values as a percentage of the predicted value, for ProbWait and ASA.

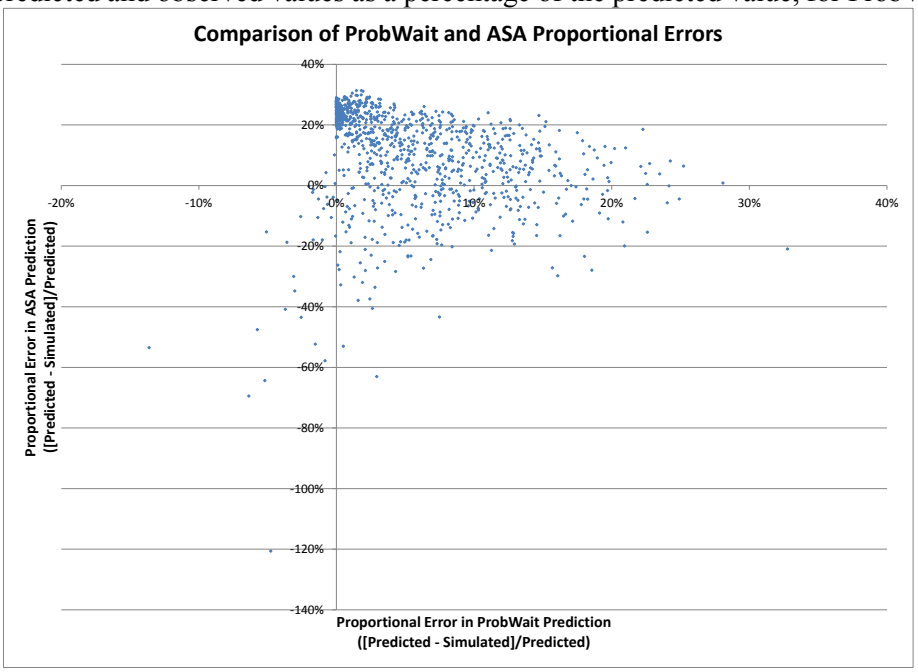

Figure 2: A comparison of the Proportional Error in ASA.

This graph demonstrates the relatively weak correlation between these errors when measured in proportional terms. As the proportional error in the ProbWait increases, the magnitude of the ASA error tends to decrease. More than 21% of the design points evaluated map to quadrant II of this graph; where the proportion of customers waiting is less than predicted, but the time to answer is more than predicted.

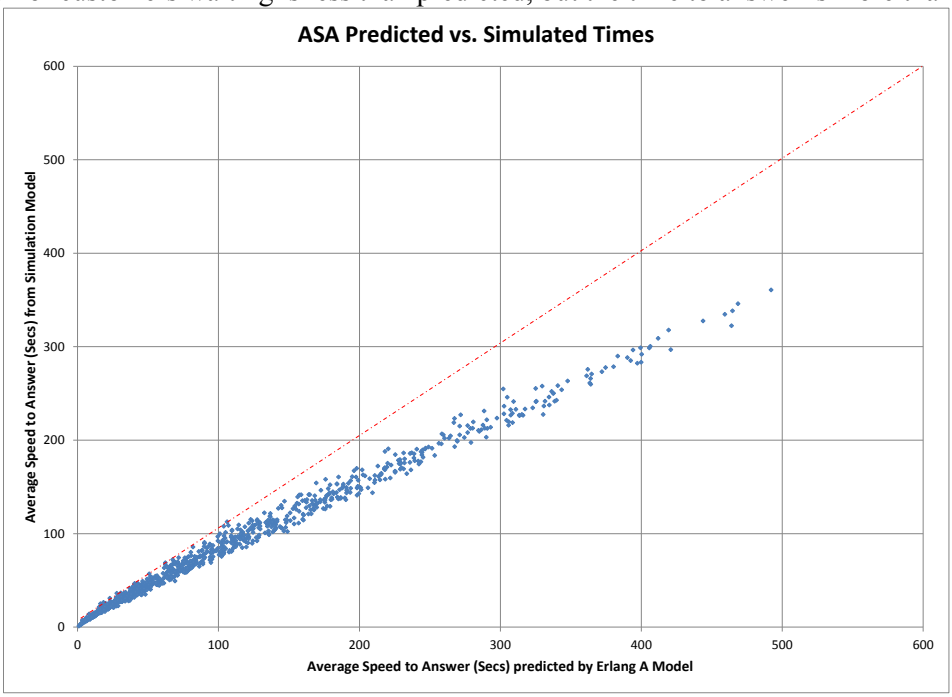

Figure 3: A scatter plot of ASA predicted and simulated values.

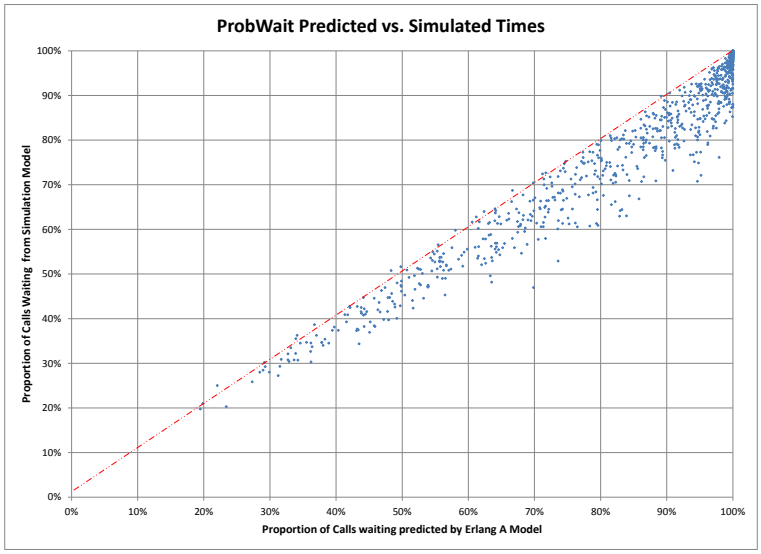

Figure 4: A scatter plot of ProbWait predicted and simulated values.

Figure **5** plots a histogram of the ProbWait error further illustrating the positive (pessimistic) bias of the ProbWait calculation. The average error is 5%, with errors ranging from -3% to 24%. The data has a strong positive skew with a sample skew measure of 1.19.

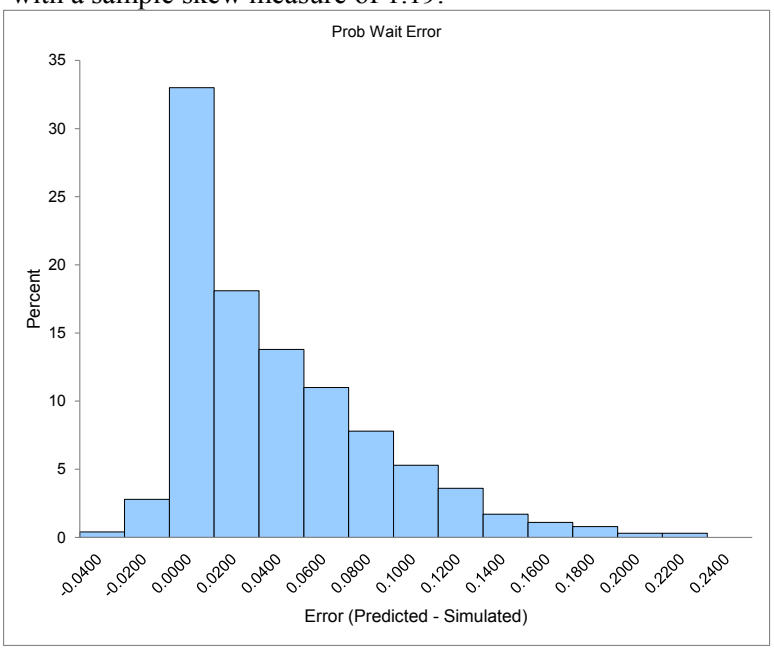

Figure 5: A histogram of ProbWait errors.

4.3 Drivers of Erlang A Errors

To better understand how each of the nine experimental factors impacts ProbWait error, we perform a regression analysis. The dependent variable is the ProbWait error. For the independent variables we use the nine experimental factors normalized to a [-1,1] scale. This normalization allows us to better assess the relative impact of each factor. The results of the regression analysis are shown in Table **3**.

Table 3: Regression Analysis of ProbWait Errors.

Regression Analysis

R^2	0.513		
Adjusted R^2	0.508	n	1000
R	0.716	k	9
Std. Error	0.032	Dep. Var.	**Prob Wait**

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	1.0570	9	0.1174	115.79	5.82E-148
Residual	1.0041	990	0.0010		
Total	2.0611	999			

Regression output					*confidence interval*	
variables	*coefficients*	*std. error*	*t (df=990)*	*p-value*	*95% lower*	*95% upper*
Intercept	0.0463	0.0010	46.007	4.38E-248	0.0444	0.0483
Num Agents	0.0089	0.0017	5.111	3.85E-07	0.0055	0.0124
Utilization Target	-0.0307	0.0018	-17.519	4.64E-60	-0.0341	-0.0272
Talk Time	-0.0055	0.0017	-3.142	.0017	-0.0089	-0.0021
Patience	0.0081	0.0017	4.663	3.54E-06	0.0047	0.0116
AR CV	0.0401	0.0018	22.928	1.17E-93	0.0367	0.0436
Talk Time CV	0.0042	0.0018	2.413	.0160	0.0008	0.0077
Patience Shape	-0.0046	0.0018	-2.626	.0088	-0.0080	-0.0012
Probability of Balking	0.0214	0.0018	12.207	5.01E-32	0.0180	0.0249
Agent Heterogeneity	0.0089	0.0017	5.105	3.96E-07	0.0055	0.0124

The model is statistically significant with a reasonably high R^2 value of .513. Each of the nine independent variables have a statistically significant effect on the error, but three factors, ARCV, utilization target, and probability of balking, have a dominant effect as illustrated by the relative magnitude of their coefficients.

Arrival rate uncertainty has the largest effect. As the arrival rate becomes less certain, the relative error in the ProbWait prediction, on average, increases in a positive (pessimistic) direction. With higher levels of uncertainty the real system tends to perform much better than predicted by the Erlang A model. It is somewhat counterintuitive that uncertain arrival rates would lead to a system with better performance than predicted. We believe this related to the s-shape of the ProbWait curve, an issue we explore in more detail below. The second largest impact comes from utilization target which has an effect in the opposite direction. Higher levels of utilization bias the error in the negative or optimistic direction. The next most pronounced effect comes from the probability of balking, which also biases the error in a positive or pessimistic direction. Closer examination reveals that these relationships may be more complicated than implied by the linear regression model. Figure **6** presents a scatter plot of the ProbWait error as a function of the offered utilization.

With relatively low offered utilization rates, the error in the ProbWait measure tends to be large, and highly variable. As the offered utilization rate increases the range of errors decreases. At relatively high loads the error varies between 0 and 7.6%. At lower utilization levels the error is as high as 24%. Of course, it is important to recall that while offered utilization may exceed 100%, the realized utilization cannot. The system equilibrates through abandonment to keep actual utilization below 100%. Figure **7** presents a scatter plot of the ProbWait error versus the realized utilization level.

This plot shows a somewhat different shape than Figure **6**. Relatively high error rates persist across most of the range of high realized utilization levels. To examine the impact of the offered utilization more directly we develop and execute several additional experiments. Figure **8** shows the results of an experiment in which we vary offered utilization and arrival rate uncertainty while holding all other factors at the midpoint. We vary offered utilization from 50% to 200% under three conditions; a known arrival rate, moderate uncertainty, and high uncertainty.

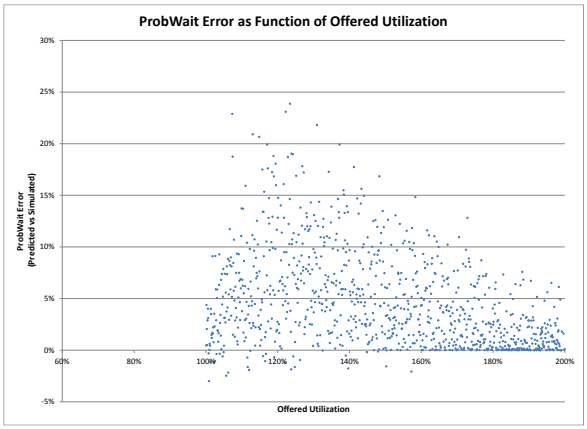

Figure 6: ProbWait Errors vs. Offered Utilization.

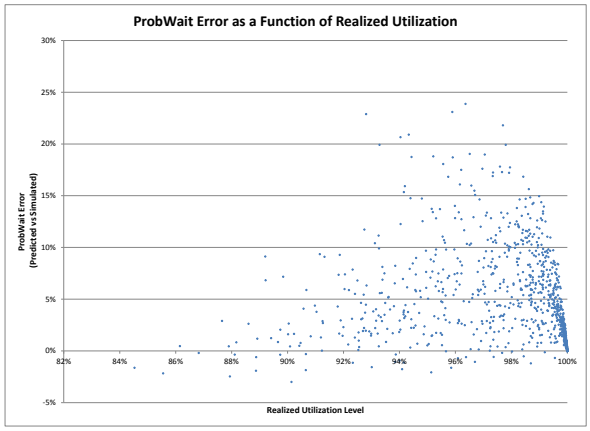

Figure 7: ProbWait Errors vs. Realized Utilization.

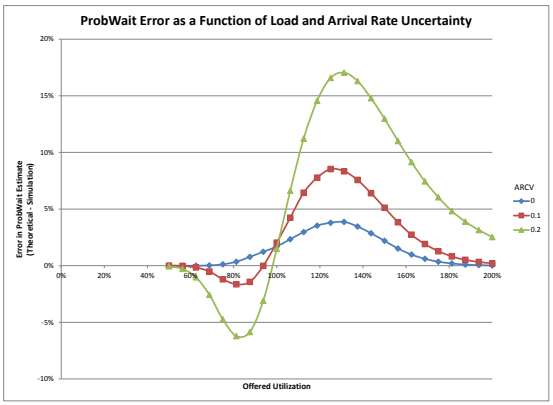

Figure 8: ProbWait Errors by AR Uncertainty.

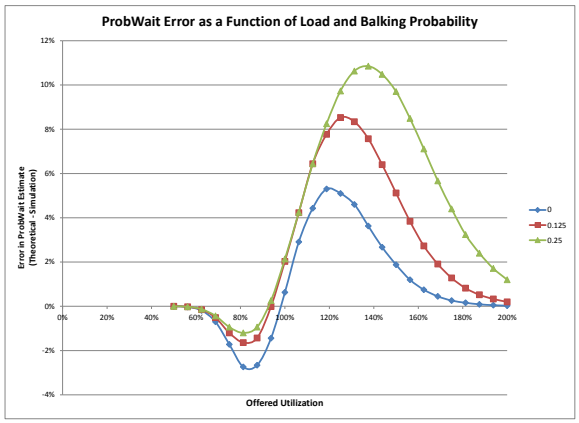

Figure 9: ProbWait Errors by Balking.

When the arrival rates are known the Erlang A model is quite accurate, in particular for relatively low or high loads. The maximum error of 4% occurs with an offered utilization level of 125%. When arrival rate uncertainty is introduced the error in the measurement becomes more pronounced, the relative maximum still occurs with utilization levels in the 125% range, but the magnitude of the error increases substantially, with errors as high as 16.5%. The graphs also reveal a region of negative bias when offered utilization levels are below 100%.

A similar phenomenon is demonstrated in another experiment illustrated in Figure **9**. In this experiment we again vary the offered utilization level, but also vary balking rate from 0 to 25%, while holding all other factors at their midpoint. Again we see very low error at the extremes, a moderate negative bias with utilization levels in the 80% range, and a strong positive bias with utilization's in the 110% to 160% range. In all cases increased balking proportion biases the error in a positive direction. Recall that balkers have no patience, they abandon as soon as they are placed in queue. Since they abandon immediately, they get out of the way and allow better results for callers who are willing to wait, improving the overall performance of the system relative to what was predicted by the Erlang A model. An implication is that actions that induce rapid abandonment, such as announcements of anticipated wait times, can improve overall system performance. To better understand the utilization effect we examine the theoretical relationship between Offered Utilization and ProbWait as predicted by the Erlang A model. Figure **10** plots the proportion of callers that must wait as a function of offered utilization when all other factors are held at their midpoint.

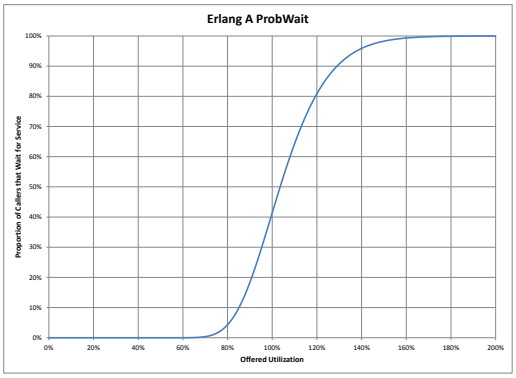

Figure 10: Erlang A Calculation of ProbWait as a Function of Offered Utilization.

Figure **10** reveals a familiar S-curve shape. With offered utilization levels in the range of 75% to 90%, the region in which we observed a negative bias, the graph has a convex shape. With offered utilization levels above about 120%, the graph becomes concave. Above 160% the graph flattens out as the probability of waiting asymptotically approaches 100%. Overall this leads to a negative bias as the improvement from lower than anticipated call volume cannot offset the degradation from higher-than-expected call volume. In the higher region of utilization, say 120% to 160%, the situation is reversed. In this region, volumes higher than anticipated have little impact as nearly 100% of callers will queue. However, volumes lower than anticipated offer substantial reduction in the proportion of calls that must wait. Since these effects don't balance out overall the prediction tends to be pessimistic, and the system on average performs with a lower proportion of callers waiting. Now recall the impact that arrival rate uncertainty has on our measurement. Instead of taking a single point measure at the anticipated arrival rate, we take the average of a sample of measures symmetrically scattered about that anticipated arrival rate. When utilization levels are relatively low, say 80%, negative deviations and arrival rate have little impact as the proportion waiting is near zero. Positive deviations on the other hand, have a significant impact dramatically increasing the relative proportion of calls that must wait. In regions of relatively high utilization this effect is reversed.

5 SUMMARY AND CONCLUSIONS

In this paper we analyzed the ability of the Erlang A model to accurately predict long-term, steady-state behavior of call centers in high traffic scenarios. In this region the Erlang A model is a reasonably good approximation of actual performance, but when actual conditions deviate from model assumptions significant error is introduced. The errors across different performance metrics are correlated in a statistically significant way, but only loosely. It is not uncommon for the errors on different metrics to be reversed. Overall our study confirms that under realistic conditions even the more sophisticated Erlang A model is subject to significant error.

6 REFERENCES

Aksin, Z., M. Armony, and V. Mehrotra. 2007. "The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research", *Production and Operations Management*, 16: 665-88.

Armony, M., and A. R. Ward. 2008. "Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems." Operations Research 58(3):624–637.

Avramidis, A. N., M. Gendreau, P. L'Ecuyer, and O. Pisacane. 2007. "Simulation-Based Optimization of Agent Scheduling in Multiskill Call Centers." In *2007 Industrial Simulation Conference*.

Bassamboo, A., J. M. Harrison, and A. Zeevi. 2005. "Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based Method", *Operations Research*, 54: 419-35.

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: 36-50.

Chen, B. P. K. , and S. G. Henderson. 2001. "Two Issues in Setting Call Centre Staffing Levels"', *Annals of Operations Research*: 175-92.

Gans, N. , G. Koole, and A. Mandelbaum. 2003. "Telephone call centers: Tutorial, Review, and Research Prospects.", *Manufacturing & Service Operations Management*, 5: 79-141.

Garnett, O , A. Mandelbaum, and M. I. Reiman. 2002. "Designing a Call Center with Impatient Customers", *Manufacturing & Service Operations Management*, 4: 208-27.

L'Ecuyer, P. 1999. "Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators", *Operations Research*, 47: 159-64.

Law, A. M. 2007. *Simulation Modeling and Analysis* (McGraw-Hill: Boston).

Mandelbaum, A., and S. Zeltyn. 2007. "Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers." In *Advances in services innovations*. 17–45. Berlin: Springer.

Mandelbaum A., Sakov A. , and Zeltyn S. 2001. "Empirical Analysis of a Call Center." Technical report. Technion - Israel Institute of Technology.

Mandelbaum, A., and Sergey Z.. 2009. "The M/M/n+G Queue: Summary of Performance Measures." Technical Note, Technion, Israel Institute of Technology.

Palm, C. 1957. "Research on Telephone Traffic Carried by Full Availability Groups", *Tele*, 1: 107.

Robbins, T. R. 2007. "Managing Service Capacity Under Uncertainty - Unpublished PhD Dissertation http://personal.ecu.edu/robbinst/ " Pennsylvania State University- Smeal College of Business.

Robbins, T. R., and T. P. Harrison. 2010. "A Stochastic Programming Model for Scheduling Call Centers with Global Service Level Agreements", *European Journal of Operational Research*, 207: 1608-19.

Robbins, T. R., D. J. Medeiros, and P. Dum. 2006. "Evaluating Arrival Rate Uncertainty in Call Centers." In *Proceedings of the 2006 Winter Simulation Conference*. edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Robbins, T., R., D. J. Medeiros, and T. P. Harrison. 2010. "Does the Erlang C Model Fit in Real Call Centers?" In *Proceedings of the 2010 Winter Simulation Conference*. edited by B. Johanssan, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2853--2864. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Santner, T. J., Brian J. Williams, and W. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

Steckley, S. G., S. G. Henderson, and V. Mehrotra. 2009. "Forecast Errors in Service Systems", *Probability in the Engineering and Informational Sciences*: 305-32.

Whitt, W. 2005. "Engineering Solution of a Basic Call-Center Model", *Management Science*, 51: 221-35.

Whitt, W. 2006a. "Fluid Models for Multiserver Queues with Abandonments", *Operations Research*, 54: 37-54.

Whitt, W. 2006b. "Sensitivity of Performance in the Erlang A Model to Changes in the Model Parameters", *Operations Research*, 54: 247-60.

Whitt, W. 2006c. "Staffing a Call Center with Uncertain Arrival Rate and Absenteeism", *Production and Operations Management*, 15: 88-102.

AUTHOR BIOGRAPHY

THOMAS R. ROBBINS is an Assistant Professor in the department of Marketing and Supply Chain at East Carolina University. He holds a PhD in Business Administration and Operations Research from Penn State University, an MBA from Case Western Reserve and a BSEE from Penn State. His email address is robbinst@ecu.edu.