Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

PERFORMANCE COMPARISON OF MSER-5 AND N-SKART ON THE SIMULATION START-UP PROBLEM

Anup C. Mokashi
Jeremy J. Tejada
Saeideh Yousefi

Edward P. Fitts Department of
Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695, U.S.A.

Tianxiang Xu
James R. Wilson

Edward P. Fitts Department of
Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695, U.S.A.

Ali Tafazzoli

Metron Aviation, Inc.
45300 Catalina Ct. Suite 101
Dulles, VA 20166, U.S.A.

Natalie M. Steiger

Maine Business School
University of Maine
Orono, ME 04469, U.S.A.

ABSTRACT

We summarize some results from an extensive performance comparison of the procedures MSER-5 and N-Skart for handling the simulation start-up problem. We assume a fixed-length simulation-generated time series from which point and confidence-interval (CI) estimators of the steady-state mean are sought. MSER-5 uses the data-truncation point that minimizes the half-length of the usual batch-means CI computed from the truncated data set. N-Skart uses a randomness test to determine the data-truncation point beyond which spaced batch means are approximately independent of each other and the simulation's initial condition; then using truncated nonspaced batch means, N-Skart exploits separate adjustments to the CI half-length that account for the effects on the distribution of the underlying Student's t-statistic arising from skewness and autocorrelation of the batch means. In most of the test problems, N-Skart's point estimator had smaller bias than that of MSER-5; moreover in all cases, N-Skart's CI estimator outperformed that of MSER-5.

1. INTRODUCTION

In many simulation studies, we are interested in estimating the characteristics of a dynamic stochastic system in steady-state operation. A steady-state (nonterminating) simulation, unlike a finite-horizon (terminating) simulation, does not have specified starting or stopping conditions. As a result, many steady-state parameters of interest can be estimated using averages accumulated over simulated time as the time horizon tends to infinity. In the execution of a steady-state simulation experiment, an arbitrary starting condition is usually chosen; and the simulation is run for a sufficient number of output responses so as to estimate the long-run average behavior of the system with a level of accuracy that is adequate for the purposes of the application at hand. Ideally, the starting condition should not affect the simulation-based statistics. However, in general the starting condition gives rise to a transient in the sequence of simulation responses that produces biased estimates of the steady-state parameters of interest. The simulation start-up problem (also known as the initialization-bias problem or the problem of the initial transient) has been the subject of many studies in the past. In this paper, we compare the performance of two recent procedures for handling the simulation start-up problem—namely, MSER-5 (White, Cobb, and Spratt 2000; Franklin and White 2008; White and Robinson 2009) and N-Skart (Tafazzoli 2009; Tafazzoli and Wilson 2009; Tafazzoli, Steiger, and Wilson 2010).

MSER-5 is a variant of the MSER procedure first proposed by White (1997). Given a finite sequence of simulation-generated observations, MSER-5 first computes batch means from adjacent (nonoverlapping) batches,

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

each consisting of five observations; then MSER-5 computes the usual batch-means point and confidence-interval (CI) estimates for the steady-state mean response in the situation that the batch means are randomly sampled from a normal distribution whose expected value coincides with the steady-state mean; and finally MSER-5 sequentially recomputes the batch-means point and CI estimators after deleting progressively more leading batch means until the half-length of the resulting CI about the grand average of the truncated batch means is minimized, subject to the constraint that the data-truncation point (that is, the length of the warm-up period) must be less than half the number of batch means. (If an even smaller CI half-length can be obtained by deleting at least the first half of the given series of batch means, then the size of the original data set is considered to be too small; and in this situation MSER-5 fails to deliver an estimator of the steady-state mean.) If MSER-5 successfully determines an acceptable truncation point, then the final truncated sequence of batch means is considered to be approximately free of initialization bias; and the corresponding grand average of the truncated batch means is supposed to have minimal mean squared error as a point estimator of the steady-state mean.

In most steady-state simulation studies, we seek not only a point estimator but also a valid CI estimator of the steady-state mean (Law 2007); and in this connection, Franklin and White (2008) made the following statement about MSER-5:

> . . . it optimizes on the objective function we most often care about in simulation studies, the confidence interval about the mean of a statistic.

Because the operation of MSER-5 is based on examining the behavior of batch-means CI estimators of the steady-state mean as a function of the truncation point, it is also natural to attempt to build a CI centered on MSER-5's final point estimator; and although we have been unable to identify any specific recommendations on how to do this in the current literature on MSER-5, we adopt the same procedure used by White and Robinson (2009)—namely, we apply the classical method of nonoverlapping batch means with 20 batches to the truncated data set finally delivered by MSER-5.

N-Skart (Tafazzoli 2009; Tafazzoli and Wilson 2009; Tafazzoli, Steiger, and Wilson 2010) is a nonsequential procedure designed to deliver a CI for the steady-state mean of a simulation output process when a single simulation-generated time series of arbitrary size is supplied by the user, and a required coverage probability for the CI is specified. N-Skart iteratively applies the randomness test of von Neumann (1941) to spaced batch means to determine sufficiently large sizes for each batch and its preceding spacer such that beyond the initial spacer (which is taken to define the data-truncation point), the spaced batch means are approximately independent of each other and the simulation's initial condition. Then using truncated, nonspaced batch means, N-Skart makes separate adjustments to the classical batch-means CI in order to account for the effects on the underlying Student's t-statistic arising from skewness and autocorrelation of the batch means. The skewness adjustment is based on a Cornish-Fisher expansion for the classical batch-means t-statistic, and the autocorrelation adjustment is based on a first-order autoregressive approximation to the batch-means autocorrelation function. If the sample size is large enough, then N-Skart delivers a valid CI for the steady-state mean as well as a point estimator that is approximately free of initialization bias. If the sample size is not large enough to yield acceptable results in the randomness-testing step, then N-Skart issues a warning message and gives the user options either to stop or to continue anyway with the computation of point and CI estimators. In this paper, we assume that N-Skart is always used with the latter option so that N-Skart automatically generates point and CI estimates of the steady-state mean for every data set to which it is applied, without any intervention by the user.

The rest of this paper is organized as follows. In Section 2 we describe in some detail the versions of MSER-5 and N-Skart that were used in the performance comparison of the two procedures. In Section 3 we describe the performance measures used in our study to evaluate the point and CI estimators delivered by MSER-5 and N-Skart; and we summarize the experimental results for a test process consisting of waiting times in the $M/M/1$ queue with an empty-and-idle initial condition and a steady-state server utilization of 90%. This paper is based on a preliminary performance evaluation by Mokashi (2010); and complete results for all test processes used in the final performance evaluation are presented in Mokashi and Wilson (2010). The slides for the oral presentation of this article are available online via `<www.ise.ncsu.edu/jwilson/files/mokashi-wsc10-pres.pdf>`.

2. OVERVIEW OF PROCEDURES TO BE COMPARED

2.1 Overview of MSER-5

MSER-5 (White, Cobb, and Spratt 2000; Franklin and White 2008; White and Robinson 2009) is a modification of the Marginal Confidence Rule (MCR) or the Marginal Standard Error Rule (MSER) proposed by White (1997). To deliver an improved simulation-based point estimator of the steady-state mean, MSER and MSER-5 aim at balancing improved accuracy achieved by reducing the estimator's bias against the loss of precision (increased variance) caused by truncating the original data set through deletion of some leading observations. Both MSER and MSER-5 are based on the following rationale: given a time series $\{X_i : i = 1,\ldots,N\}$ of simulation-generated responses having fixed length (sample size) N from which we seek to compute an accurate estimator of the steady-state mean $\mu_X = \lim_{i\to\infty} \mathrm{E}[X_i]$, we seek to find a data-truncation point beyond which all the remaining observations are typical of steady-state behavior. For each candidate data-truncation point in the data set, we compute a CI for μ_X based on all the observations beyond the truncation point; and we take the half-length of this CI as a measure of the extent to which all the remaining observations are typical of steady-state behavior, where a smaller CI half-length indicates closer conformity to steady-state behavior. It follows that the data-truncation truncation point should be set to minimize the length of the CI for μ_X based on the remaining (truncated) output sequence.

The main differences between MSER and MSER-5 are the following: (a) Whereas MCR works directly with the individual simulation-generated observations $\{X_i : i = 1,\ldots,N\}$, MSER-5 operates on nonoverlapping batch means with batch size 5 in order to ensure more stable behavior in the CIs used to determine the truncation point; and (b) MSER-5 constrains the data-truncation point to be in the first half of the given data set. Therefore, MSER5 uses as its basic data items the batch means with batch size 5,

$$Z_j = \frac{1}{5}\sum_{i=1}^{5} X_{5(j-1)+i} \quad \text{for} \quad j = 1,\ldots,k = \lfloor N/5 \rfloor\,,$$

where for each real number u the floor function $\lfloor u \rfloor$ denotes the greatest integer not exceeding u.

If d denotes the data-truncation point (that is, the length of the warm-up period), then the grand average and sample variance of the truncated batch means $\{Z_j : j = d+1,\ldots,k\}$ are

$$\overline{Z}(k,d) = \frac{1}{k-d}\sum_{j=d+1}^{k} Z_j \quad \text{and} \quad S_Z^2(k,d) = \frac{1}{k-d}\sum_{j=d+1}^{k}\left[Z_j - \overline{Z}(k,d)\right]^2, \tag{1}$$

respectively; and the naive $100(1-\alpha)\%$ CI for μ_X based on (1) has the form

$$\overline{Z}(k,d) \pm z_{1-\alpha/2}\frac{S_Z(k,d)}{\sqrt{k-d}}\,, \tag{2}$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution. White, Cobb, and Spratt (2000) do not recommend using (2) as the final CI estimator for μ_X; instead they recommend merely using (2) as a device for determining the optimal truncation point d^* as follows:

$$d^* = \mathop{\arg\min}_{0\le d\le\lfloor k/2\rfloor} z_{1-\alpha/2}\frac{S_Z(k,d)}{\sqrt{k-d}}\text{; but if } d^* = \lfloor k/2\rfloor\text{, then MSER-5 fails because of inadequate sample size.} \tag{3}$$

Unfortunately when $d^* = \lfloor k/2 \rfloor$ so that MSER-5 fails to deliver any estimator of μ_X, White, Cobb, and Spratt (2000) do not suggest a method for increasing the sample size so that MSER-5 can ultimately deliver the desired point and CI estimators of μ_X.

If $d^* < \lfloor k/2 \rfloor$ in (3), then MSER-5 delivers the truncated sample mean $\overline{Z}(k,d^*)$ as the final point estimator of the steady-state mean μ_X. To compute the associated nominal $100(1-\alpha)\%$ CI for μ_X, we follow the approach used by White and Robinson (2009). To be specific, we apply the classical method of nonoverlapping batch means to the truncated sequence $\{Z_j : j = d^*+1,\ldots,k\}$, which is now regarded as the "original" (raw, unbatched) observations from which we compute

$$k^* = 20 \text{ “new” batch means with batch size } m^* = \lfloor (k - d^*)/k^* \rfloor .$$

Therefore the ℓth new batch mean is computed as

$$\overline{Z}_\ell(m^*, d^*) = \frac{1}{m^*} \sum_{j=1}^{m^*} Z_{d^* + (\ell-1)m^* + j} \quad \text{for} \quad \ell = 1, \ldots, k^* ;$$

and the corresponding grand average and sample variance of the new batch means are given by

$$\overline{\overline{Z}}(k^*, m^*, d^*) = \frac{1}{k^*} \sum_{\ell=1}^{k^*} \overline{Z}_\ell(m^*, d^*) \quad \text{and} \quad S^2_{\overline{Z}}(k^*, m^*, d^*) = \frac{1}{k^* - 1} \sum_{\ell=1}^{k^*} \Big[\overline{Z}_\ell(m^*, d^*) - \overline{\overline{Z}}(k^*, m^*, d^*) \Big]^2 ,$$

respectively. In terms of the statistics $\overline{Z}(k, d^*)$ and $S_{\overline{Z}}(k^*, m^*, d^*)$), the final $100(1-\alpha)\%$ CI for μ_X is

$$\overline{Z}(k, d^*) \pm t_{1-\alpha/2, k^*-1} \frac{S_{\overline{Z}}(k^*, m^*, d^*)}{\sqrt{k^*}} , \tag{4}$$

where $t_{1-\alpha/2, k^*-1}$ denotes the $1-\alpha/2$ quantile of Student's t-distribution with k^*-1 degrees of freedom. Although it is not clear that White and Robinson (2009) recommend the CI estimator (4) for general use in conjunction with MSER-5, we use (4) at least as a straw man intended to stimulate the development of other CI estimators that are specifically designed for use with MSER-5. In any case, the CI (4) is consistent with the recommendations of Schmeiser (1982) on applying the method of nonoverlapping batch means to a data set $\{Z_j : j = d^* + 1, \ldots, k\}$ that is a realization of a covariance stationary simulation output process.

2.2 Overview of N-Skart

The input to N-Skart is a simulation-generated time series $\{X_j : i = 1, \ldots, N\}$ of fixed length N, where $N \geq 1{,}280$; and the user specifies the required coverage probability $1-\alpha$ (where $0 < \alpha < 1$) for a CI estimator of μ_X based on the given data set. N-Skart handles the start-up problem by applying the randomness test of von Neumann (1941) to determine sufficiently large values of the batch size m and spacer size dm (where $m \geq 1$ and $d \geq 0$) such that the corresponding k spaced batch means

$$Y_j(m, d) = \frac{1}{m} \sum_{i=1}^{m} X_{\{j(d+1)-1\}m+i} \quad \text{for} \quad j = 1, \ldots, k \tag{5}$$

are approximately independent of each other and of the initial condition X_0. Because the spacer preceding the jth batch of size m consists of the ignored (deleted) observations $\{X_i : i = (j-1)(d+1)m + 1, \ldots, [j(d+1)-1]m\}$, we see that the first spacer ($j = 1$) consists of the observations $\{X_i : i = 1, \ldots, dm\}$ so that the first spaced batch mean $Y_1(m, d) = m^{-1} \sum_{i=1}^{m} X_{dm+i}$ is approximately independent of the initial condition X_0; moreover all the spaced batch means $\{Y_j(m, d) : j = 1, \ldots, k\}$ are approximately independent of each other and the initial condition X_0. It follows that any effects due to initialization bias are limited to the initial spacer $\{X_i : i = 1, \ldots, dm\}$; and this is the reason why N-Skart uses the initial spacer as the warm-up period so that the first dm observations are deleted (ignored). If data set size N is not large enough to enable N-Skart to determine sufficiently large values for the spacer size and batch size such that the spaced batch means pass the randomness test, then N-Skart issues a warning and gives the user options either to stop or to continue anyway in computing point and CI estimators of μ_X.

Beyond the truncation point dm, N-Skart computes k' truncated, nonspaced batch means with batch size m,

$$Y_j(m) = \frac{1}{m} \sum_{i=1}^{m} X_{(d+j-1)m+i} \quad \text{for} \quad j = 1, \ldots, k' ,$$

where k' is taken large enough to use as much of the data set $\{X_i : i = 1,\ldots,N\}$ as possible; and then N-Skart computes the sample mean and variance of the truncated, nonspaced batch means,

$$\overline{Y}(m,k') = \frac{1}{k'}\sum_{j=1}^{k'} Y_j(m) \quad \text{and} \quad S^2_{m,k'} = \frac{1}{k'-1}\sum_{j=1}^{k'}\left[Y_j(m) - \overline{Y}(m,k')\right]^2,$$

respectively. Finally N-Skart delivers an asymptotically valid $100(1-\alpha)\%$ skewness- and autocorrelation-adjusted CI for μ_X having the form

$$\left[\overline{Y}(m,k') - G(L)\sqrt{AS^2_{m,k'}/k'},\ \overline{Y}(m,k') - G(R)\sqrt{AS^2_{m,k'}/k'}\right], \tag{6}$$

where the skewness adjustments $G(L)$ and $G(R)$ are defined in terms of the function

$$G(\zeta) = \left[\sqrt[3]{1+6\beta(\zeta-\beta)} - 1\right]\Big/(2\beta) \ \text{ for all real } \zeta \text{, where } \quad \beta = \widehat{\mathscr{B}}_{m,k''}\Big/\left(6\sqrt{k''}\right), \tag{7}$$

and

$$\widehat{\mathscr{B}}_{m,k''} = \left\{\begin{array}{l}\text{approximately unbiased estimator of the marginal skewness of } Y_j(m) \text{ computed from the}\\ k'' \text{ spaced batch means of the form (5) with batch size } m \text{ that are separated by spacers}\\ \text{of size at least } dm\text{, where } k'' \text{ is taken large enough to use the entire data set of size } N\end{array}\right\}$$

so that skewness-adjustment function $G(\cdot)$ has the arguments

$$L = t_{1-\alpha/2,k''-1} \quad \text{and} \quad R = t_{\alpha/2,k''-1}$$

where for $q \in (0,1)$, the quantity $t_{q,\nu}$ denotes the q quantile of Student's t-distribution with ν degrees of freedom; and the correlation adjustment A is computed as

$$A = \left[1 + \widehat{\varphi}_{Y(m)}\right]\Big/\left[1 - \widehat{\varphi}_{Y(m)}\right],$$

where the standard estimator of the lag-one correlation of the truncated, nonspaced batch means (2.2) is

$$\widehat{\varphi}_{Y(m)} = \frac{1}{k'-1}\sum_{j=1}^{k'-1}\frac{[Y_j(m) - \overline{Y}(m,k')][Y_{j+1}(m) - \overline{Y}(m,k')]}{S^2_{m,k'}}.$$

(Note that in (7), the indicated cube root $\sqrt[3]{1+6\beta(\zeta-\beta)}$ is understood to have the same sign as the quantity $1+6\beta(\zeta-\beta)$.) The specific methods for computing m, d, k', k'', and $\widehat{\mathscr{B}}_{m,k''}$ are explained in Tafazzoli, Steiger, and Wilson (2010).

3. PERFORMANCE COMPARISON OF N-SKART AND MSER-5

Many different approaches have been proposed to evaluate the effects of initial conditions on the performance of point and CI estimators for μ_X. In the preliminary performance evaluation of Mokashi (2010), several test processes were chosen to be representative of the level of complexity observed in large-scale simulation applications. Some other test processes were selected that exhibit extreme stochastic behavior, and these cases were used as stress tests for MSER-5 and N-Skart. Each test process has a steady-state mean μ_X that either can be obtained analytically or can be evaluated numerically to a high degree of accuracy; and therefore we could compare the performance of MSER-5 and N-Skart with respect to the accuracy and reliability of their point and CI estimators of μ_X. In this article we summarize the results of the final performance evaluation presented in Mokashi and Wilson (2010) for one test process that is a nearly universal benchmark for evaluating steady-state simulation analysis procedures—namely, queue waiting times in the $M/M/1$ queue with an empty-and-idle initial condition and a steady-state server utilization of 90%.

3.1 Performance Measures

The effectiveness of MSER-5 and N-Skart in removing the initial transient can be measured in terms of the bias, variance, and mean squared error of the point estimator of μ_X delivered by each method. The bias measures systematic deviation of the point estimator away from the true steady-state mean μ_X, while the variance measures the random variation around the point estimator's expected value. Truncation methods require considerably smaller sample sizes to reduce the initial-transient effects in the simulation output compared with other methods for handling the simulation start-up problem. However, this may result in a significant increase in the variance of the truncated sample mean. The mean squared error is a standard measure of the accuracy of the truncated sample mean as an estimator of μ_X that combines both the bias and variance of this estimator,

$$\mathrm{MSE}\big[\overline{Y}(m,k')\big] \equiv \mathrm{E}\big[\big\{\overline{Y}(m,k') - \mu_X\big\}^2\big] = \mathrm{Bias}^2\big[\overline{Y}(m,k')\big] + \mathrm{Var}\big[\overline{Y}(m,k')\big]\,, \tag{8}$$

where

$$\mathrm{Bias}\big[\overline{Y}(m,k')\big] \equiv \mathrm{E}\big[\overline{Y}(m,k')\big] - \mu_X \quad \text{and} \quad \mathrm{Var}\big[\overline{Y}(m,k')\big] \equiv \mathrm{E}\big[\big\{\overline{Y}(m,k') - \mathrm{E}\big[\overline{Y}(m,k')\big]\big\}^2\big] \tag{9}$$

when we are evaluating the performance of the truncated sample mean $\overline{Y}(m,k')$ delivered by N-Skart; and equations similar to (8) and (9) apply to the truncated sample mean $\overline{Z}(k,d^*)$ delivered by MSER-5. We will use the bias, variance and mean squared error as the main performance measures for comparison of the point estimators of μ_X delivered by MSER-5 and N-Skart.

The effectiveness of the truncation methods in estimating μ_X can also be measured in terms of the following properties of CIs computed with each procedure using nominal coverage probabilities (confidence coefficients) of 90% and 95%:

- CI coverage is the probability that the steady-state mean μ_X falls within the CI;
- CI relative precision is the ratio of the CI half-length to the magnitude of the corresponding point estimator (usually the CI's midpoint);
- CI half-length measures the precision of the CI estimator; and
- Variance of the CI half-length measures the variability of the CI estimator.

In the case of N-Skart, the "half-length" of N-Skart's CI (6) is taken to be $\max\{\,|G(L)|, |G(R)|\,\}\sqrt{AS^2_{m,k'}/k'}$, the maximum of the left- and right-hand subintervals of (6) with respect to the point estimator $\overline{Y}(m,k')$.

In our study, we generated 1,000 replications of each test process; and we compared the performance of MSER-5 and N-Skart for samples of size N = 10,000, 20,000, 50,000 and 200,000. These sample sizes were selected in an attempt to characterize the performance of MSER-5 and N-Skart in "small," "medium," and "large" data sets. We applied both MSER-5 and N-Skart to the same realizations of each test process to sharpen the performance comparison of the two procedures. Over 1,000 independent replications of each test process with the four selected sample sizes, we computed the following average performance measures for N-Skart and MSER-5:

- Empirical CI coverage probability;
- Average CI relative precision;
- Average CI half-length;
- Variance of the CI half-length;
- Bias of the truncated sample mean;
- Variance of the truncated sample mean; and
- Mean squared error of the truncated sample mean.

In addition to these performance measures, for each scenario described above we also examined a histogram depicting the empirical frequency distribution of the truncated sample mean delivered by MSER-5 and N-Skart.

For MSER-5, the above performance measures were conditional given a successful result in applying the procedure; and therefore the empirical performance measures were averaged over all replications for which MSER-5 successfully delivered point and CI estimators of μ_X. In addition, we reported the "unconditional CI coverage" delivered by MSER-5, which is the percentage of all 1,000 replications on which MSER-5 delivered a CI that covered the steady-state mean μ_X. This latter statistic is intended to characterize the performance of the CI (4) that can be expected on a single application of MSER-5 in practice. For N-Skart, we selected the option to deliver point and CI estimators of μ_X regardless of whether or not the randomness-testing step of the procedure was completed successfully; and thus the performance measures reported for N-Skart are typical of completely automated use of the procedure with no intervention by the user. Our method for reporting the performance of MSER-5 and N-Skart is consistent with all the previously cited papers on both procedures.

3.2 Results for M/M/1 Queue-Waiting-Time Process with Empty Initial Condition and 90% Utilization

In this section we summarize the results obtained in the $M/M/1$ queue-waiting-time process $\{X_i : i = 1, \ldots, N\}$ in which X_i denotes the waiting time in the queue for the ith customer, where N = 10,000, 20,000, 50,000, and 200,000. The interarrival times for the customers are randomly sampled from an exponential distribution with mean $1/\lambda$ and corresponding arrival rate of $\lambda = 0.9$ customers per unit time, and the service times for customers are randomly sampled from an exponential distribution with mean $1/\mu$ and corresponding service rate of $\mu = 1.0$ customers per unit time. Thus, the steady-state server utilization for this system is $\rho = \lambda/\mu = 0.9$. The system starts in the empty-and-idle state so that $X_1 = 0$ on every replication of the process $\{X_i\}$. The steady-state expected waiting time in the queue is $\mu_X = \rho/[\mu(1-\rho)] = 9.0$ time units.

The $M/M/1$ queue-waiting-time process is characterized by a relatively short warm-up period. However, the process exhibits a pronounced autocorrelation structure, with the autocorrelation function for the waiting time decaying slowly as the lag increases. Also, the $M/M/1$ queue waiting times have a steady-state probability distribution which has a nonzero probability mass at zero and a exponential tail. This results in a slow convergence of the batch means to the normal distribution with increasing batch size.

Table 1 summarizes the performance of MSER-5 and N-Skart on the selected $M/M/1$ queue-waiting-time process. From the results in Table 1, it is evident that the CI properties obtained from N-Skart were better than those obtained from MSER-5. The CI coverages delivered by N-Skart were close to the corresponding nominal coverage levels. For smaller sample sizes, N-Skart delivered CI coverages slightly below the nominal level; but N-Skart also delivered a CI with large relative precision, indicating that a larger sample size was required in order to have practically useful CIs. For example with the sample size N = 10,000, the nominal 95% CIs delivered by N-Skart had an empirical coverage probability of 92.9% and an average relative precision of 57.1%. As the sample size increased, the CI coverage delivered by N-Skart was close to the nominal coverage level; and in most cases, the actual coverage was slightly larger than the nominal CI coverage. For example with the sample size N = 200,000, the nominal 95% CIs delivered by N-Skart had an empirical coverage probability of 96.3% and an average relative precision of 11.8%. In contrast, the CIs delivered by MSER-5 exhibited empirical CI coverages that were significantly below the corresponding nominal levels. For example with the sample size N = 10,000, the nominal 95% CIs delivered by MSER-5 had a conditional empirical coverage probability of 75.6%, given that MSER-5 successfully delivered point and CI estimators of μ_X; but our estimate of the unconditional probability that MSER-5 will deliver a CI covering the steady-state mean μ_X was only 57.1%. For the sample sizes 10,000, 20,000, and 50,000, we concluded that the performance of N-Skart's CIs was substantially better than that of MSER-5. For the sample size 200,000, we concluded that both methods delivered acceptable CIs.

With respect to the point estimators of μ_X delivered by MSER-5 and N-Skart, we observed that for all sample sizes considered, the point-estimator bias was substantially larger for MSER-5 than for N-Skart. For the sample sizes 10,000 and 20,000, the variance of the point estimator delivered by N-Skart was larger than that for MSER-5. This suggests that at least on some runs, N-Skart truncated at least half the sample data, while MSER-5 either truncated less than half the sample data or simply failed to deliver a truncation point. For all sample sizes considered, the point-estimator bias was an order of magnitude smaller for N-Skart than for MSER-5.

The distributions of the truncated sample means delivered by MSER-5 and N-Skart are depicted in Figures 1–4. For the sample size 10,000, the distribution of MSER-5's truncated sample mean $\overline{Z}(k, d^*)$ was clearly shifted to the left of the steady-state mean μ_X.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1 Conclusions

Considering the results delivered by N-Skart for all the test processes, we observed that N-Skart delivered estimates of the steady-state mean whose values were usually centered close to the steady-state mean, as is evident from

Table 1: Performance of MSER-5 and N-Skart in the $M/M/1$ queue-waiting-time process with 90% server utilization and empty-and-idle initial condition.

Results for MSER-5 Algorithm (1)–(4)

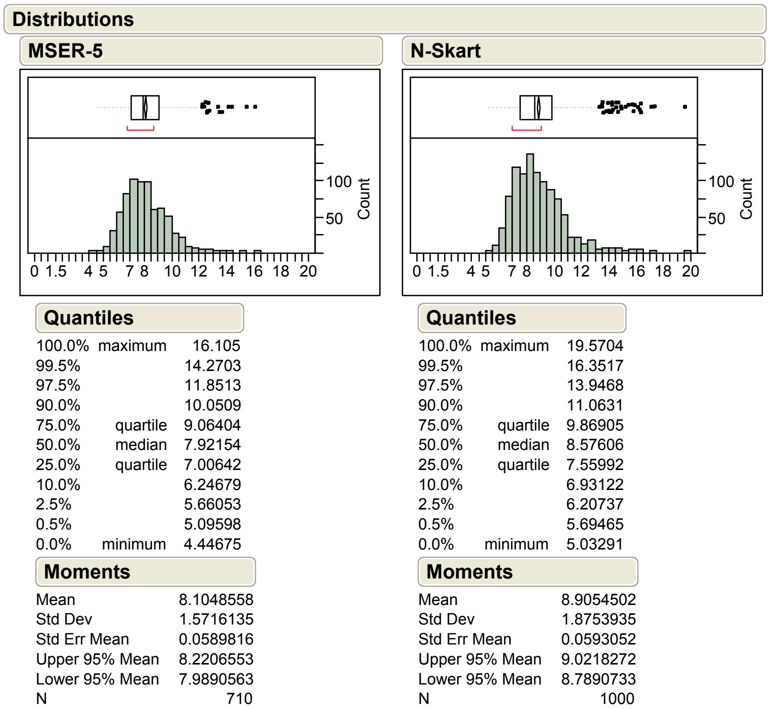
Confidence-Interval Properties		Overall Sample Size N			
$1-\alpha$	Empirical Perf. Meas.†	10,000	20,000	50,000	200,000
90%	Uncond. CI coverage	48.70%	60.60%	73.40%	83.10%
	CI coverage	68.59%	77.89%	84.76%	88.31%
	Avg. rel. prec.	27.03%	21.78%	14.96%	7.84%
	Avg. CI half-length	2.17943	1.8689	1.3173	0.70474
	Var. CI half-length	1.1235	0.75447	0.26442	0.040649
95%	Uncond. CI coverage	53.70%	66.70%	76.80%	88.90%
	CI Coverage	75.63%	85.73%	88.68%	94.47%
	Avg. rel. prec.	32.72%	26.36%	18.12%	9.49%
	Avg. CI half-length	2.63825	2.2623	1.5947	0.85311
	Var. CI half-length	1.6464	1.1056	0.38748	0.059567
Empirical Point-Estimator		Overall Sample Size N			
Performance Measures		10,000	20,000	50,000	200,000
Trunc. Sample Mean		8.1049	8.5383	8.7597	8.9563
MSE		3.2678	1.5966	0.69154	0.17049
Variance		2.4665	1.3834	0.63378	0.16858
|Bias|		0.89514	0.46171	0.24035	0.043725
# Failures in 1,000 Runs		290	222	134	59

†Unconditional CI coverage is averaged over all 1,000 runs; all other empirical performance measures for MSER-5 are conditional given success of the procedure and thus are averaged over all successful runs of MSER-5.

Results for N-Skart Algorithm (5)–(7)

Confidence-Interval Properties		Overall Sample Size N			
$1-\alpha$	Empirical Perf. Meas.	10,000	20,000	50,000	200,000
90%	CI coverage	87.70%	91.80%	92.50%	92.10%
	Avg. rel. prec.	43.07%	32.50%	20.71%	9.51%
	Avg. CI half-length	4.0592	3.0394	1.8932	0.86168
	Var. CI half-length	10.350	9.2895	1.4728	0.074006
95%	CI Coverage	92.90%	95.70%	95.50%	96.30%
	Avg. rel. prec.	57.05%	42.81%	26.93%	11.75%
	Avg. CI half-length	5.3664	4.0040	2.4653	1.0655
	Var. CI half-length	19.244	17.733	3.1065	0.18158
Empirical Point-Estimator		Overall Sample Size N			
Performance Measures		10,000	20,000	50,000	200,000
Trunc. Sample Mean		8.9055	8.9445	8.9625	9.0062
MSE		3.5225	1.6052	0.65584	0.17491
Variance		3.5136	1.6022	0.65444	0.17487
|Bias|		0.09455	0.055517	0.037485	0.0061757

Figures 1–4 presented in the previous section. The empirical CI coverage probabilities were in close conformance with the user-specified coverage levels. It was observed that for test cases with small sample sizes, N-Skart still produced valid CI estimates albeit with large values for the relative precision of the CI ($>$ 40%). Such a large value for the CI's relative precision is usually an indication that because of substantial bias, correlation, or nonnormality in the target process (or some combination of these anomalous characteristics), it is necessary to increase the sample size in order to obtain practically useful results. In general to determine a sample size that is sufficiently large to ensure reliable performance of N-Skart, we recommend performing a pilot study in which Skart, the fully sequential variant of N-Skart, is applied to an initial sample whose size is practically feasible for the application at hand (Tafazzoli and Wilson 2010). In such a pilot study, the application of Skart to the initial sample will either deliver the desired CI or return an estimate of the size N of the sample that should be collected and supplied to N-Skart. So long as the relative precision of the CI delivered by N-Skart does not exceed 40%, then in all our computational



Distributions

MSER-5

Quantiles

100.0%	maximum	16.105
99.5%		14.2703
97.5%		11.8513
90.0%		10.0509
75.0%	quartile	9.06404
50.0%	median	7.92154
25.0%	quartile	7.00642
10.0%		6.24679
2.5%		5.66053
0.5%		5.09598
0.0%	minimum	4.44675

Moments

Mean	8.1048558
Std Dev	1.5716135
Std Err Mean	0.0589816
Upper 95% Mean	8.2206553
Lower 95% Mean	7.9890563
N	710

N-Skart

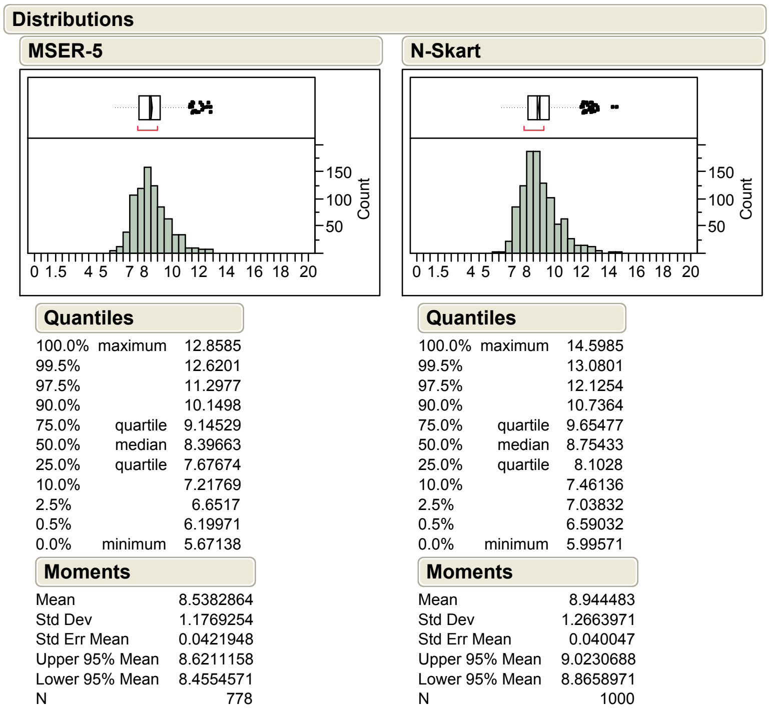
Quantiles

100.0%	maximum	19.5704
99.5%		16.3517
97.5%		13.9468
90.0%		11.0631
75.0%	quartile	9.86905
50.0%	median	8.57606
25.0%	quartile	7.55992
10.0%		6.93122
2.5%		6.20737
0.5%		5.69465
0.0%	minimum	5.03291

Moments

Mean	8.9054502
Std Dev	1.8753935
Std Err Mean	0.0593052
Upper 95% Mean	9.0218272
Lower 95% Mean	8.7890733
N	1000

Figure 1: Empirical distributions of truncated sample mean for MSER-5 and N-Skart when applied to $M/M/1$ queue-waiting-time process with $X_1 = 0$, $\rho = 0.9$, and $N = 10,000$.



Distributions

MSER-5

Quantiles

100.0%	maximum	12.8585
99.5%		12.6201
97.5%		11.2977
90.0%		10.1498
75.0%	quartile	9.14529
50.0%	median	8.39663
25.0%	quartile	7.67674
10.0%		7.21769
2.5%		6.6517
0.5%		6.19971
0.0%	minimum	5.67138

Moments

Mean	8.5382864
Std Dev	1.1769254
Std Err Mean	0.0421948
Upper 95% Mean	8.6211158
Lower 95% Mean	8.4554571
N	778

N-Skart

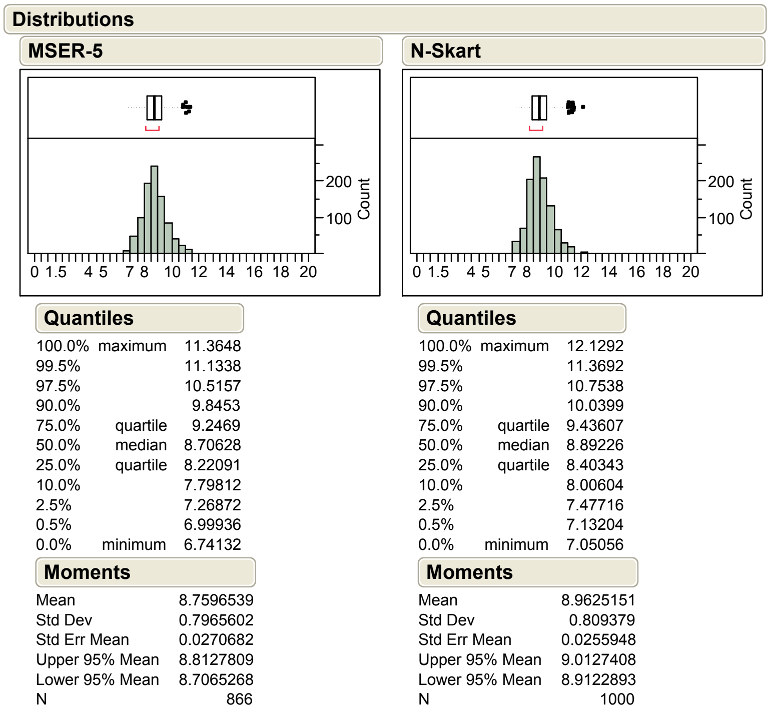
Quantiles

100.0%	maximum	14.5985
99.5%		13.0801
97.5%		12.1254
90.0%		10.7364
75.0%	quartile	9.65477
50.0%	median	8.75433
25.0%	quartile	8.1028
10.0%		7.46136
2.5%		7.03832
0.5%		6.59032
0.0%	minimum	5.99571

Moments

Mean	8.944483
Std Dev	1.2663971
Std Err Mean	0.040047
Upper 95% Mean	9.0230688
Lower 95% Mean	8.8658971
N	1000

Figure 2: Empirical distributions of truncated sample mean for N-Skart and MSER-5 when applied to $M/M/1$ queue-waiting-time process with $X_1 = 0$, $\rho = 0.9$, and $N = 20,000$.

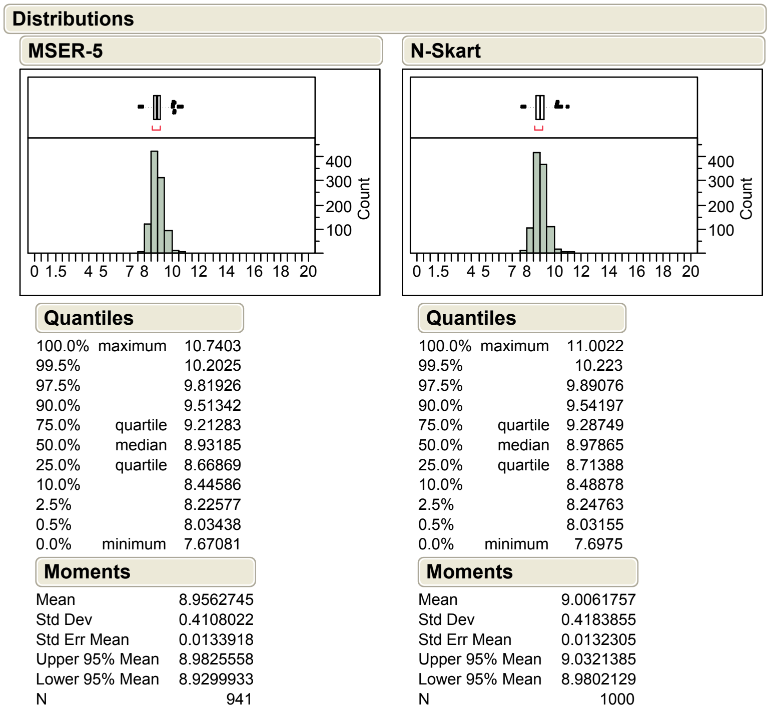


```
Distributions
MSER-5                                  N-Skart

Quantiles                               Quantiles
100.0%  maximum   11.3648               100.0%  maximum   12.1292
99.5%             11.1338               99.5%             11.3692
97.5%             10.5157               97.5%             10.7538
90.0%              9.8453               90.0%             10.0399
75.0%    quartile  9.2469               75.0%    quartile  9.43607
50.0%    median    8.70628              50.0%    median    8.89226
25.0%    quartile  8.22091              25.0%    quartile  8.40343
10.0%              7.79812              10.0%              8.00604
2.5%               7.26872              2.5%               7.47716
0.5%               6.99936              0.5%               7.13204
0.0%     minimum   6.74132              0.0%     minimum   7.05056

Moments                                 Moments
Mean              8.7596539             Mean              8.9625151
Std Dev           0.7965602             Std Dev            0.809379
Std Err Mean      0.0270682             Std Err Mean      0.0255948
Upper 95% Mean    8.8127809             Upper 95% Mean    9.0127408
Lower 95% Mean    8.7065268             Lower 95% Mean    8.9122893
N                       866             N                      1000
```

Figure 3: Empirical distributions of truncated sample mean for MSER-5 and N-Skart when applied to $M/M/1$ queue-waiting-time process with $X_1 = 0$, $\rho = 0.9$, and $N = 50,000$.



```
Distributions
MSER-5                                  N-Skart

Quantiles                               Quantiles
100.0%  maximum   10.7403               100.0%  maximum   11.0022
99.5%             10.2025               99.5%             10.223
97.5%              9.81926              97.5%              9.89076
90.0%              9.51342              90.0%              9.54197
75.0%    quartile  9.21283              75.0%    quartile  9.28749
50.0%    median    8.93185              50.0%    median    8.97865
25.0%    quartile  8.66869              25.0%    quartile  8.71388
10.0%              8.44586              10.0%              8.48878
2.5%               8.22577              2.5%               8.24763
0.5%               8.03438              0.5%               8.03155
0.0%     minimum   7.67081              0.0%     minimum    7.6975

Moments                                 Moments
Mean              8.9562745             Mean              9.0061757
Std Dev           0.4108022             Std Dev           0.4183855
Std Err Mean      0.0133918             Std Err Mean      0.0132305
Upper 95% Mean    8.9825558             Upper 95% Mean    9.0321385
Lower 95% Mean    8.9299933             Lower 95% Mean    8.9802129
N                       941             N                      1000
```

Figure 4: Empirical distributions of truncated sample mean for MSER-5 and N-Skart when applied to $M/M/1$ queue-waiting-time process with $X_1 = 0$, $\rho = 0.9$, and $N = 200,000$.

experience both point and CI estimates of the steady-state mean are approximately free of initialization bias; and the actual coverage probability of the CI will be fairly close to the nominal coverage level.

Given a simulation output response of arbitrary length, the MSER-5 truncation heuristic is designed to estimate a truncation point such that the truncated sequence is approximately free of initialization bias. The optimum MSER-5 test statistic minimizes the width of the CI centered on the truncated sample mean. This method is appealing intuitively and is much simpler to implement in practice compared with N-Skart. However, from the results in the previous section, we concluded that the point estimates for the steady-state mean provided by MSER-5 exhibited considerable bias, especially for smaller sample sizes. Also, the CI coverages delivered by MSER-5 did not conform to the user-specified coverage levels for the sample sizes 10,000, 20,000, and 50,000. As mentioned in Section 2.1, the CI (4) is at best a straw man intended to stimulate the development of other CI estimators that are specifically designed for use with MSER-5.

Thus, considering the above discussion regarding the individual performances of N-Skart and MSER-5, we concluded that N-Skart outperformed MSER-5 in the $M/M/1$ queue-waiting time process with empty-and-idle initial condition and 90% server utilization. Similar results were obtained for a wide range of other test processes, as detailed in Mokashi (2010) and Mokashi and Wilson (2010).

4.2 Recommendations for Future Work

On the basis of this research, the following recommendations have been made for future work:

- The main function of the MSER-5 heuristic is to deliver a truncated sample mean based on the truncation point estimated by the MSER-5 test statistic. Since it is desirable to have a valid CI associated with the point estimator of the steady-state mean, it would be helpful to combine the MSER-5 heuristic with a procedure that delivers a valid CI estimator.
- The MSER-5 heuristic is a special case of the MSER-*m* heuristic which divides the given simulation output sequence into batches of size *m*. A fixed batch size limits the flexibility of this procedure to account for different degrees of correlation in various stochastic processes. It would be highly desirable to augment MSER-*m* with an automatic procedure for determining an appropriate value of the batch size *m* on each application of the procedure.
- Implementing N-Skart is much more difficult that implementing MSER-5. Simplified, computationally efficient versions of N-Skart should be implemented in portable, robust software that can be easily invoked “on the fly” in standard simulation environments or on a stand-alone basis.
- The experimental performance evaluation of Mokashi and Wilson (2010) should be substantially expanded to include a much greater diversity of test processes with different types of transient behavior.

REFERENCES

Franklin, W. W., and K. P. White, Jr. 2008. Stationarity tests and MSER-5: Exploring the intuition behind mean-squared-error reduction in detecting and correcting initialization bias. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 541–546. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-sim.org/wsc08papers/064.pdf> [accessed July 16, 2010].

Law, A. M. 2007. *Simulation modeling and analysis*. 4th ed. New York: McGraw-Hill, Inc.

Mokashi, A. C. 2010. The simulation start-up problem: Performance comparison of N-Skart and MSER-5. Master’s thesis, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina. Available online via <www.ise.ncsu.edu/jwilson/files/mokashi10ms.pdf> [accessed July 16, 2010].

Mokashi, A. C., and J. R. Wilson. 2010. The simulation start-up problem: Performance comparison of N-Skart and MSER-5. Technical Report, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina. Available online via <www.ise.ncsu.edu/jwilson/files/mokashi10tr.pdf> [accessed July 16, 2010].

Schmeiser, B. W. 1982. Batch size effects in the analysis of simulation output. *Operations Research* 30:556-568.

Tafazzoli, A. 2009. *Skart: A skewness- and autoregression-adjusted batch-means procedure for simulation analysis*. Ph.D. thesis, Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina. Available via <www.lib.ncsu.edu/resolver/1840.16/3868> [accessed July 16, 2010].

Tafazzoli, A., N. M. Steiger, and J. R. Wilson. 2010a. N-Skart: A nonsequential skewness- and autoregression-adjusted batch-means procedure for simulation analysis. *IEEE Transactions on Automatic Control* forthcoming. Preprint available online via <www.ise.ncsu.edu/jwilson/files/tafazzoli10ieeetac.pdf> [accessed July 16, 2010].

Tafazzoli, A., and J. R. Wilson. 2009. N-Skart: A nonsequential skewness- and autoregression-adjusted batch-means procedure for simulation analysis. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 652–662. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online as www.informs-sim.org/wsc09papers/063.pdf [accessed July 16, 2010].

Tafazzoli, A., and J. R. Wilson. 2010. Skart: A skewness- and autoregression-adjusted batch-means procedure for simulation analysis. *IIE Transactions* forthcoming. Preprint available online via <www.ise.ncsu.edu/jwilson/files/tafazzoli10iiet.pdf> [accessed July 16, 2010].

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics* 12:367–395.

White, K. P., M. J. Cobb, and S. C. Spratt. 2000. A comparison of five steady-state truncation heuristics for simulation. In *Proceedings of the 2000 Winter Simulation Conference*, ed. R. R. Barton, J. A. Joines, P. A. Fishwick, and K. Kang, 755–760. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online as <www.informs-sim.org/wsc00papers/099.PDF> [accessed July 16, 2010].

White, K. P., and S. Robinson. 2009. The problem of the initial transient (again), or why MSER works. In *Proceedings of the 2009 INFORMS Simulation Society Research Workshop*, ed. L. H. Lee, M. E. Kuhl, J. W. Fowler, and S. Robinson, 90–95. Baltimore: Institute for Operations Research and the Management Sciences. Available online as <www.informs-sim.org/2009informs-simworkshop/paper92-97.pdf> [accessed July 16, 2010].

AUTHOR BIOGRAPHIES

ANUP C. MOKASHI is an M.S. Candidate in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of IIE and INFORMS. His e-mail address is <acmokash@ncsu.edu>.

JEREMY J. TEJADA is a Ph.D. Candidate in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of IIE and INFORMS. His e-mail address is <jjtejada@ncsu.edu>.

SAEIDEH YOUSEFI is an M.S. Candidate in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. She is a member of INFORMS. Her e-mail address is <syousef@ncsu.edu>.

TIANXIANG XU is an undergraduate in the Department of Control Science and Engineering at Zhejiang University in China. In the summer of 2010, he participated in the Summer Research Program at North Carolina State University. His e-mail address is <shawnxtx@gmail.com>.

JAMES R. WILSON is professor of the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of ACM and SCS, and he is a Fellow of IIE and INFORMS. His e-mail address is <jwilson@ncsu.edu>, and his web page is <www.ise.ncsu.edu/jwilson>.

ALI TAFAZZOLI is a senior analyst at Metron Aviation, Inc. He is a member of IIE and INFORMS. His e-mail address is <tafazzoli@metronaviation.com>.

NATALIE M. STEIGER is an associate professor of production and operations management in the University of Maine Business School. She is a member of IIE and INFORMS. Her e-mail address is <nsteiger@maine.edu>.