Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

REVISITING SUBSET SELECTION

David J. Eckman
Matthew Plumlee
Barry L. Nelson

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA

ABSTRACT

In the subset-selection approach to ranking and selection, a decision-maker seeks a subset of simulated systems that contains the best with high probability. We present a new, generalized framework for constructing these subsets and demonstrate that some existing subset-selection procedures are situated within this framework. The subsets are built by calculating, for each system, a minimum standardized discrepancy between the observed performances and the space of problem instances for which that system is the best. A system's minimum standardized discrepancy is then compared to a cutoff to determine whether the system is included in the subset. We examine the problem of finding the tightest statistically valid cutoff for each system and draw connections between our approach and other subset-selection methodologies. Simulation experiments demonstrate how the screening power and subset size are affected by the choice of standardized discrepancy.

1 INTRODUCTION

We consider the classical ranking-and-selection (R&S) problem of selecting the simulated alternative (system) with the best expected performance from among a finite set. A well-studied approach to this problem involves returning a subset of systems to the decision-maker, referred to as *subset selection* (Gupta 1965; Alam and Rizvi 1966). Instead of selecting a single system as the best, subset selection affords the decision-maker the opportunity to make a final selection based on secondary considerations, which may not be captured by the simulation model. The decision-maker can also infer from the size of the subset whether there are multiple near-optimal systems. As the trend towards ubiquitous parallel computing gives rise to R&S problems with thousands or even millions of systems, subset-selection procedures are likely to only grow in importance.

Commercial simulation software like Simio and Arena® allows users to easily employ subset-selection procedures in a variety of ways as part of their experiments. Subset selection can be run as a stand-alone procedure, taking a fixed number of replications from each system and reporting a final subset. Subset selection can also be used to "clean-up" after running a search (Boesel et al. 2003) or to efficiently screen out inferior systems before running a more intensive simulation-optimization algorithm (Nelson et al. 2001). In the latter case, some second-stage R&S procedures reuse replications taken by a first-stage subset-selection procedure and deliver an overall statistical guarantee on the selected system; the additive decomposition lemma of Nelson et al. (2001) is an illustrative example.

The development of subset-selection procedures has been somewhat ad hoc, featuring a variety of goals and assumptions. Many goals are expressed in terms of a class of solutions which should be included in the returned subset with some specified probability. Examples include retaining the optimal system, one or all δ-optimal systems, or the m best systems—in terms of expected performance—with high probability. Arguably, the most popular of these goals is the probability of correct selection (PCS) guarantee, which

states that the subset contains the best system—or one tagged as the best, if there are ties—with sufficiently high probability for any problem instance. A closely related goal is to guarantee that the expected false elimination rate is sufficiently small (Pei et al. 2018).

Gupta (1965) developed the first subset-selection procedure delivering the PCS guarantee, under an assumption of a known, common variance of systems' outputs and a common sample size. Subsequent procedures relaxed these assumptions to handle unknown, unequal variances and unequal sample sizes (Boesel et al. 2003). Procedures such as these construct the subset based on pairwise comparisons; that is, the estimated performance of a given system is compared to that of each other system. The PCS guarantee can be assured by carefully choosing the allowed difference between the estimated performances of paired systems. In contrast to this approach, we introduce a broader subset-selection paradigm for delivering the PCS guarantee that encompasses pairwise methods. Our approach provides a new lens through which we illuminate how subset selection works as well as a rich space within which subset-selection procedures can be developed and analyzed.

Our framework entails computing an index for each system and comparing it to a cutoff—one chosen to deliver the PCS guarantee—to determine whether to include that system in the subset. Borrowing ideas from isotonic regression (Barlow et al. 1972; Robertson et al. 1988; Silvapulle and Sen 2005), we use as an index the minimum standardized discrepancy between the estimated performances and the space of problem instances for which that system is one of the best. Plumlee and Nelson (2018) proposed this discrepancy-based model for a different setting in which only a subset of systems are simulated, and the decision-maker possesses information relating the expected performances of some systems. Apropos of that line of research, this paper eliminates the functional information component, yet generalizes the standardized discrepancies and cutoffs. We also provide deeper insights into the framework's application to subset selection and draw connections to existing methods.

2 SUBSET-SELECTION GUARANTEES

Suppose there are k systems under consideration and for each System i, n_i replications are obtained, where $n_1,\ldots,n_k$ are fixed in advance. Simulating System i produces outputs $X_{i1},X_{i2},\ldots,X_{in_i}$ satisfying the following distributional assumption:

Assumption 1 For $i=1,\ldots,k$, the outputs $X_{i1},X_{i2},\ldots,X_{in_i}$ are independent and normally distributed with unknown mean μ_i and known variance σ_i^2. Furthermore, outputs from different systems are independent.

The normality component of Assumption 1 is standard in the R&S literature and can be approximately satisfied by batching outputs and applying the Central Limit Theorem. It allows us to devise specific procedures that deliver finite-sample statistical guarantees. The other aspects of Assumption 1 are made solely for ease of presentation; e.g., the proposed methods can be adapted to address the use of common random numbers (CRN) across systems. We make the (unrealistic) assumption of known variances to more clearly introduce the main ideas and form connections to other subset-selection procedures. Our methods can also be modified to handle unknown variances; see Plumlee and Nelson (2018) for a preliminary treatment and Section 6 for further remarks.

Under Assumption 1, the vector $\mu=(\mu_1,\mu_2,\ldots,\mu_k)$ uniquely describes the sampling distribution of the outputs, hence we refer to μ as the problem instance. Given the outputs, μ is estimated by the vector of sample means, $\widehat{\mu}$, where $\widehat{\mu}_i={n_i}^{-1}\sum_{l=1}^{n_i}X_{il}$ for $i=1,\ldots,k$. We use subscript $[\cdot]$ to indicate the (unknown) ordering of the expected performances, i.e., $\mu_{[1]}\leq\mu_{[2]}\leq\cdots\leq\mu_{[k]}$, and we assume without loss of generality that a larger expected performance is better—thus System $[k]$ is (one of) the best. If multiple systems are tied for the best, we assume that one of them is arbitrarily tagged as $[k]$.

A subset-selection procedure returning a subset $S\subseteq\{1,2,\ldots,k\}$ delivers the PCS guarantee if

$$\text{PCS}\equiv\mathbb{P}([k]\in S)\geq 1-\alpha \text{ for all } \mu\in\mathbb{R}^k, \tag{1}$$

where $1-\alpha$ is specified by the decision-maker and $\mathbb{P}$ denotes the probability measure associated with taking n_i replications from System i for $i=1,\ldots,k$, according to Assumption 1. Since the system tagged

as $[k]$ is chosen arbitrarily from among those tied for the best, the PCS guarantee ensures that for each such system, the probability it is included in S exceeds $1-\alpha$. The PCS guarantee is trivially delivered by a procedure that always returns $S = \{1,2,\ldots,k\}$, however uninformative this may be to the decision-maker. Our interest is therefore in subset-selection procedures that deliver the PCS guarantee while having a small expected subset size.

The PCS guarantee is unattainable without knowledge of the family of distributions of the outputs, or some form of control on their tail behavior, e.g., bounds. Assumption 1, however, allows us to design subset-selection procedures that deliver it. If the outputs are not normally distributed, then under suitable conditions, the proposed methods achieve the PCS guarantee asymptotically (as $\min_{i=1,\ldots,k} n_i \to \infty$) via the Central Limit Theorem.

3 A DISCREPANCY-BASED FRAMEWORK

We present a discrepancy-based framework within which subset-selection procedures can be designed to deliver the PCS guarantee.

3.1 Overview

The central idea of our approach is to calculate an index for each system and include in the subset all systems whose indices are below corresponding cutoffs. A given system's index is intended to reflect how well the sample data agrees with the hypothesis that that system is one of the best, with smaller values indicating closer agreement. To mathematically describe this relationship, we introduce the *standardized discrepancy* between $\widehat{\mu}$ and a vector of performances $m = (m_1,\ldots,m_k)$, expressed as $d(m,\widehat{\mu})$. The standardized discrepancy is also implicitly a function of the sampling variances, $\sigma_1^2/n_1,\ldots,\sigma_k^2/n_k$, which play the role of standardizing differences between components of m and $\widehat{\mu}$.

For each system, we obtain an index by minimizing $d(m,\widehat{\mu})$ subject to the constraint that that system is the best according to m. Let $\mathsf{M}_i \equiv \{m : m_i \geq m_j \text{ for all } j \neq i\}$ denote the set of performance vectors for which System i is one of the best. The *minimum standardized discrepancy* (i.e., index) of System i is defined as

$$D_i(\widehat{\mu}) \equiv \min_{m \in \mathsf{M}_i} d(m,\widehat{\mu}). \tag{2}$$

The index $D_i(\widehat{\mu})$ can be interpreted as the minimum distance between $\widehat{\mu}$ and M_i. Smaller values of $D_i(\widehat{\mu})$ signify that the sample data more strongly supports the claim that System i is one of the best. As a function of $\widehat{\mu}$, the index $D_i(\widehat{\mu})$ is a random variable whose distribution depends on the unknown problem instance.

For an arbitrary System i, we consider a deterministic cutoff, denoted by D_i, satisfying

$$\mathbb{P}(D_i(\widehat{\mu}) \leq \mathsf{D}_i) \geq 1-\alpha \text{ for all } \mu \in \mathsf{M}_i. \tag{3}$$

In words, Condition (3) states that for any problem instance in which System i is one of the best, its index will be less than its cutoff with high probability. Let the subset S comprise the solutions for which their index is less than their cutoff, i.e.,

$$S = \{i : D_i(\widehat{\mu}) \leq \mathsf{D}_i\}.$$

PCS guarantee (1) follows as a direct consequence of choosing the cutoffs so that Condition (3) holds for all $i = 1,\ldots,k$.

3.2 Examples of Standardized Discrepancies

As examples of standardized discrepancies, we consider weighted variations of the ℓ_1, ℓ_2 and ℓ_∞ distance functions:

$$d^1(m,\widehat{\mu}) \equiv \sum_{j=1}^{k} \frac{\sqrt{n_j}}{\sigma_j}|\widehat{\mu}_j - m_j|, \quad d^2(m,\widehat{\mu}) \equiv \sum_{j=1}^{k} \frac{n_j}{\sigma_j^2}(\widehat{\mu}_j - m_j)^2, \quad \text{and} \quad d^\infty(m,\widehat{\mu}) \equiv \max_{j=1,\ldots,k} \frac{\sqrt{n_j}}{\sigma_j}|\widehat{\mu}_j - m_j|.$$

The standardization of each component-wise difference $\widehat{\mu}_j - m_j$ adjusts for the uncertainty about μ_j, as measured by the standard error of $\widehat{\mu}_j$. Consequently, for $m = \mu$, these standardized discrepancies are functions of the pivotal statistics $(\widehat{\mu}_j - \mu_j)/(\sigma_j/\sqrt{n_j})$ which have a standard normal distribution. When the variances are unknown, there exist analogs of these standardized discrepancies in which the true variances are replaced with the sample variances; see Plumlee and Nelson (2018) for an analysis of the unknown-variances analog of $d^2(m,\widehat{\mu})$.

One can also define standardized discrepancies in a way that imitates subset-selection procedures built around pairwise differences, e.g.,

$$d^{\mathrm{P}}(m,\widehat{\mu}) \equiv \max_{j,\ell=1,\ldots,k} \frac{|\widehat{\mu}_j - \widehat{\mu}_\ell - m_j + m_\ell|}{\sqrt{\sigma_j^2/n_j + \sigma_\ell^2/n_\ell}}.$$

Compared to d^∞, d^{P} scales the *differences* of differences of (j,ℓ) pairs of components of m and $\widehat{\mu}$ by the standard error of $\widehat{\mu}_j - \widehat{\mu}_\ell$. Changing a single component of m consequently affects $k-1$ terms in the maximum for $d^{\mathrm{P}}(m,\widehat{\mu})$, but only a single term in the maximum for $d^\infty(m,\widehat{\mu})$.

Computing the minimum standardized discrepancy of a given system involves solving the k-dimensional optimization problem given in Definition (2). For a fixed $\widehat{\mu}$, the standardized discrepancies above are all convex functions of m, allowing us to exploit the Karush-Kuhn-Tucker conditions and duality (Boyd and Vandenberghe 2004) to reduce the calculations. The simplified forms of the various minimum standardized discrepancies—stated in Proposition 1—entail either optimizing a one-dimensional convex function or enumerating a discrete set of cardinality k.

Proposition 1 For $i = 1,\ldots,k$,

$$D_i^1(\widehat{\mu}) \equiv \min_{m\in\mathsf{M}_i} d^1(m,\widehat{\mu}) = \min_{\overline{m}\in\mathbb{R}} \sum_{j\neq i} \frac{\sqrt{n_j}}{\sigma_j} \left[\widehat{\mu}_j - \overline{m}\right]^+ + \frac{\sqrt{n_i}}{\sigma_i} \left|\widehat{\mu}_i - \overline{m}\right|, \tag{4}$$

$$D_i^2(\widehat{\mu}) \equiv \min_{m\in\mathsf{M}_i} d^2(m,\widehat{\mu}) = \min_{\overline{m}\in\mathbb{R}} \sum_{j\neq i} \frac{n_j}{\sigma_j^2} \left(\left[\widehat{\mu}_j - \overline{m}\right]^+\right)^2 + \frac{n_i}{\sigma_i^2}(\widehat{\mu}_i - \overline{m})^2, \tag{5}$$

$$D_i^\infty(\widehat{\mu}) \equiv \min_{m\in\mathsf{M}_i} d^\infty(m,\widehat{\mu}) = \max_{j=1,\ldots,k} \frac{\widehat{\mu}_j - \widehat{\mu}_i}{\sigma_j/\sqrt{n_j} + \sigma_i/\sqrt{n_i}}, \quad \text{and} \tag{6}$$

$$D_i^{\mathrm{P}}(\widehat{\mu}) \equiv \min_{m\in\mathsf{M}_i} d^{\mathrm{P}}(m,\widehat{\mu}) = \max_{j=1,\ldots,k} \frac{\widehat{\mu}_j - \widehat{\mu}_i}{\sqrt{\sigma_j^2/n_j + \sigma_i^2/n_i}}. \tag{7}$$

Equations (4) and (5) can be easily solved: Equation (4) by enumerating each $\widehat{\mu}_j$ and Equation (5) by computing the derivative of the objective function with respect to $\overline{m}$ at each $\widehat{\mu}_j$ and finding the minimizer within the interval in which the derivative changes sign. In like manner, Equations (6) and (7) can be computed by enumerating over $j = 1,\ldots,k$. Equations (6) and (7) also show that $D_i^\infty(\widehat{\mu})$ and $D_i^{\mathrm{P}}(\widehat{\mu})$ differ only in the denominators of their constituent terms. In particular, $D_i^\infty(\widehat{\mu})$ standardizes by the sum of the standard errors of $\widehat{\mu}_i$ and $\widehat{\mu}_j$ whereas $D_i^{\mathrm{P}}(\widehat{\mu})$ standardizes by the standard error of their difference.

These four discrepancy-based methods always return a non-empty subset, since the index of the system with the largest sample mean is always zero (take $m = \widehat{\mu}$) and the cutoffs must be nonnegative to satisfy Condition (3). We discuss details of cutoffs in Section 6.

3.3 Useful Properties of Standardized Discrepancies

For the upcoming results, we require that the standardized discrepancy satisfy the following conditions, both of which are satisfied by d^1, d^2, d^∞, and d^{P}.

(C1) *Shift Invariance:* For any $c \in \mathbb{R}$, $d(m,\widehat{\mu}) = d(m + c\mathbf{1}_k, \widehat{\mu} + c\mathbf{1}_k)$ where $\mathbf{1}_k$ is a k-vector of ones.

(C2) *Monotonicity:* $D_i(\widehat{\mu})$ is monotone decreasing in $\widehat{\mu}_i$ and monotone increasing in $\widehat{\mu}_j$ for all $j \neq i$.

Condition (C1) states that if all of the sample means were shifted by the same amount, shifting all components of m likewise would yield the same standardized discrepancy. It is intended to align with the shift invariance of the normal distribution stipulated by Assumption 1. Condition (C2) articulates the intuition that, ceteris paribus, observing sample data for which a given System i looks better or another System j looks worse should strengthen our belief that System i is one of the best, i.e., decrease its index.

A closely related property to Condition (C2) is that if a given system looks worse than another in terms of estimated performance, while having less uncertainty about its expected performance, then it should have a larger index. Proposition 2 affirms that this property is satisfied by the standardized discrepancies we have considered.

Proposition 2 For d^1, d^2, d^∞, and d^{P} and any Systems i and j, if $\widehat{\mu}_i \leq \widehat{\mu}_j$ and $\sigma_i^2/n_i \leq \sigma_j^2/n_j$, then $D_i(\widehat{\mu}) \geq D_j(\widehat{\mu})$.

If the same cutoff were used for all systems, we could exploit Proposition 2 to more efficiently construct the subset, without having to compute the indices for all systems. When a given system is added to the subset, we can also include all systems with larger estimated performances and larger sampling variances. Conversely, when a given system is left out of the subset, we can screen out all systems with smaller estimated performances and smaller sampling variances.

4 CONNECTIONS

The discrepancy-based framework has several notable connections to existing subset-selection procedures and other statistical methodologies.

4.1 Gupta's Procedure and Extended Screen-to-the-Best

As previously mentioned, the classic procedure of Gupta (1965) delivers the PCS guarantee under the assumption of a common, known variance, σ^2, and a common sample size, n. The subset returned by Gupta's procedure is defined as

$$S^{\mathrm{Gupta}} \equiv \left\{ i : \widehat{\mu}_i \geq \widehat{\mu}_j - W_{ij} \text{ for all } j \neq i \right\} \quad \text{where} \quad W_{ij} = h\sqrt{2}\sigma/\sqrt{n},$$

and h is the $1-\alpha$ quantile of the maximum of $k-1$ standard normal random variables with common pairwise correlations of $1/2$ (Kim and Nelson 2006). By adding in the trivially satisfied inequality $\widehat{\mu}_i \geq \widehat{\mu}_i - W_{ii}$ and rearranging terms, we obtain

$$S^{\mathrm{Gupta}} = \left\{ i : \widehat{\mu}_j - \widehat{\mu}_i \leq W_{ij} \text{ for all } j = 1,\ldots,k \right\} = \left\{ i : \max_{j=1,\ldots,k} \frac{\widehat{\mu}_j - \widehat{\mu}_i}{\sqrt{\sigma^2/n + \sigma^2/n}} \leq h \right\}. \tag{8}$$

By substitution from Equation (7), $S^{\mathrm{Gupta}} = \{ i : D_i^{\mathrm{P}}(\widehat{\mu}) \leq h \}$. Gupta's procedure is therefore a special case of the discrepancy-based approach with standardized discrepancy d^{P} and cutoff $\mathsf{D}_i = h$ for all $i = 1,\ldots,k$. Further dividing both sides of the inequality in Equation (8) by $\sqrt{2}$ shows that Gupta's procedure also corresponds to the choice of standardized discrepancy d^∞ and cutoff $\mathsf{D}_i = h/\sqrt{2}$ for all $i = 1,\ldots,k$.

Another well-known subset-selection procedure that delivers the PCS guarantee using pairwise comparisons is the Extended Screen-to-the-Best (ESTTB) procedure of Boesel et al. (2003). The procedure is designed to handle unknown, unequal variances and unequal sample sizes. For a known-variances version of the procedure, the returned subset is defined as

$$S^{\mathrm{ESTTB}} \equiv \left\{ i : \widehat{\mu}_i \geq \widehat{\mu}_j - W_{ij} \text{ for all } j \neq i \right\} \quad \text{where} \quad W_{ij} = z_\beta \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}},$$

and z_β is the β quantile of the standard normal distribution and $\beta = (1-\alpha)^{1/(k-1)}$. Similar manipulations of the inequalities show that $S^{\text{ESTTB}} = \{i : D_i^{\text{P}}(\widehat{\mu}) \leq z_\beta\}$. The known-variances version of the ESTTB procedure is therefore a special case of the discrepancy-based approach with standardized discrepancy d^{P} and cutoff $\mathsf{D}_i = z_\beta$ for all $i = 1, \ldots, k$.

4.2 Bayesian Subset Selection

The subset-selection approach to R&S has alternatively been studied from a Bayesian perspective using various loss functions (Miescke 1979; Hamilton et al. 2008). Under the Bayesian interpretation, the decision-maker places a prior distribution on the problem instance and, after taking replications, applies the standard Bayesian updating to obtain a posterior distribution on the problem instance. The *posterior* probability of correct selection of System i, denoted by pPCS_i, is the probability—under the posterior distribution—that System i is one of the best, i.e., the posterior probability that $\mu \in \mathsf{M}_i$. One can construct a subset S^{Bayes} for which the posterior probability that S^{Bayes} includes at least one of the optimal systems exceeds $1-\alpha$ as follows:

1. For Systems $i = 1, \ldots, k$, calculate pPCS_i.
2. Sort systems in descending order by pPCS_i.
3. Add systems to S^{Bayes} until the sum of the pPCS_i terms for $i \in S^{\text{Bayes}}$ first exceeds $1-\alpha$.

Provided the posterior distribution on μ has a density, the posterior probability that an arbitrary subset $A \subseteq \{1, \ldots, k\}$ includes one of the optimal systems is exactly $\sum_{i \in A} \text{pPCS}_i$. From this property, it can be shown that the algorithm above produces the smallest subset that satisfies the desired probability statement.

The discrepancy-based and Bayesian methods for subset selection both assign an index to each system and form subsets accordingly. For a given System i, the indices $D_i^2(\widehat{\mu})$ and pPCS_i have much in common: Both are functions of the sample means and sampling variances that satisfy Pareto relationships with respect to these quantities (e.g., Proposition 2 herein and Proposition 4 of Eckman 2019). Furthermore, both terms are related to

$$f(m; \widehat{\mu}) = \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n_j}{\sigma_j^2}} \exp\left(-\frac{(m_j - \widehat{\mu}_j)^2}{2\sigma_j^2/n_j} \right),$$

when regarded as a function of m. The function $d^2(\cdot, \widehat{\mu})$ equals $-2\log f(\cdot; \widehat{\mu})$ plus a constant, therefore for System i, the minimizer of $d^2(m, \widehat{\mu})$ for $m \in \mathsf{M}_i$ is the same as the maximizer of $f(m; \widehat{\mu})$ for $m \in \mathsf{M}_i$. On the other hand, $f(m; \widehat{\mu})$ is the density of a multivariate normal distribution with mean vector $\widehat{\mu}$ and covariance matrix $\Sigma = \text{diag}(\sigma_1^2/n_1, \ldots, \sigma_k^2/n_k)$. Under the conjugate reference prior with independent beliefs, $f(m; \widehat{\mu})$ is precisely the posterior distribution of μ having observed $\widehat{\mu}$ (DeGroot 2004). In summary, the Bayesian subset-selection approach evaluates pPCS_i by integrating the density $f(m, \widehat{\mu})$ over M_i, while the discrepancy-based approach evaluates $D_i^2(\widehat{\mu})$ by effectively maximizing $f(m, \widehat{\mu})$ over M_i.

4.3 Isotonic Regression and Hypothesis Tests

Using the d^2 standardized discrepancy for subset selection involves minimizing a weighted sum of squares subject to order restrictions, a problem referred to as *isotonic regression* within the statistics community (Silvapulle and Sen 2005). In particular, $m^* = \arg\min_{m \in \mathsf{M}_i} d^2(m, \widehat{\mu})$ is called the isotonic regression of $\widehat{\mu}$ with weights $n_1/\sigma_1^2, \ldots, n_k/\sigma_k^2$, subject to the order restrictions describing System i as one of the best.

The d^2 standardized discrepancy is also closely related to testing the null hypothesis that $\mu_1 = \cdots = \mu_k$ against the alternative hypothesis that $\mu \in \mathsf{M}_i$. In particular, the index $D_i^2(\widehat{\mu})$ resembles the standardized residual sum of squares under the alternative hypothesis,

$$\text{RSS} \equiv \min_{m \in \mathsf{M}_i} \sum_{j=1}^{k} \frac{n_j}{\sigma_j^2} \sum_{l=1}^{n_j} (X_{jl} - m_j)^2,$$

which appears in various test statistics. It can be shown that m^* is also the minimizer in RSS; Section 3.2.1 of Silvapulle and Sen (2005) provides a full derivation based on writing out the loglikelihood for the outputs under Assumption 1 and making use of the sufficiency of $\widehat{\mu}$. It follows from the same argument that m^* is also the maximum likelihood estimator of μ under the alternative hypothesis.

5 COMPARATIVE ANALYSIS

We compare the four discrepancy-based methods and the Bayesian subset-selection method to better understand how they form subsets and how effective they are at screening out inferior systems.

5.1 Acceptance Regions

We consider a simple example with $k = 3$ systems to illustrate how the choice of standardized discrepancy leads to different subsets. The sampling variances are set as $\sigma_1^2/n_1 = 1$, $\sigma_2^2/n_2 = 2$, and $\sigma_3^2/n_3 = 3$, and $1-\alpha$ is set as 0.95. We fixate on System 1 and determine the values of $\widehat{\mu}$ for which System 1 would be included in the subset. The resulting regions can be interpreted as valid acceptance regions for a null hypothesis that $\mu \in \mathsf{M}_1$ at a significance level $\alpha = 0.05$.

By exploiting the shift invariance of the normal distribution, it is possible to plot these acceptance regions in $\mathbb{R}^2$ with $\widehat{\mu}_1 - \widehat{\mu}_2$ and $\widehat{\mu}_1 - \widehat{\mu}_3$ along the axes, as in Figure 1. The upper-right orthant corresponds to sample data for which System 1 looks the best, i.e., M_1. Hence for all four standardized discrepancies, $D_1(\widehat{\mu}) = 0$ in this region. In Figure 1, the acceptance regions are the areas up and to the right of the plotted curves. The boundaries correspond to the contours of $D_1(\cdot)$ at which $D_1(\widehat{\mu}) = \mathsf{D}_1$, where D_1 is the smallest cutoff satisfying Condition (3), discussed in greater detail in Section 6. The geometry of the standardized discrepancies manifests in the boundaries in the lower-left orthant: piecewise-affine for d^1, curved for d^2, and rectangular for d^∞ and d^{P}. Although similar in shape, the acceptance regions for d^∞ and d^{P} differ slightly due to the unequal sampling variances. In the upper-left and lower-right orthants, where $\widehat{\mu}_2 \geq \widehat{\mu}_1 \geq \widehat{\mu}_3$ and $\widehat{\mu}_3 \geq \widehat{\mu}_1 \geq \widehat{\mu}_2$, respectively, the boundaries are parallel to the axes for these four acceptance regions. This follows from Equations (4)–(7) in which for any system j such that $\widehat{\mu}_j \leq \widehat{\mu}_i$, reducing $\widehat{\mu}_j$ does not change the index of System i.

We similarly analyze the Bayesian subset-selection method described in Section 4.2, determining the values of $\widehat{\mu}$ for which System 1 is included in the subset. As plotted in Figure 1, the acceptance region for the Bayesian subset-selection method somewhat resembles that of d^2, with a curved boundary in the lower-left orthant. Other portions of the boundary, however, are jagged due to the discrete construction of S^{Bayes} as a function of pPCS_1, pPCS_2, and pPCS_3. The boundary comprises five curves, each of which describes certain values of $\widehat{\mu}$ for which a small perturbation in the estimated performances would change the composition of the subset. For instance, one of the curves corresponds to $\mathrm{pPCS}_1 = \alpha$, $\mathrm{pPCS}_2 \in [\alpha, 1-2\alpha]$, and $\mathrm{pPCS}_3 \in [\alpha, 1-2\alpha]$, while another corresponds to $\mathrm{pPCS}_3 = 1-\alpha$ and $\mathrm{pPCS}_1 \geq \mathrm{pPCS}_2$. It can also be seen from the nonconvex shape of the acceptance region that the indicator function of the event $\{\text{System } 1 \in S^{\mathrm{Bayes}}\}$ is not monotone in the sample means, as is the case for the discrepancy-based methods.

5.2 Inclusion Probabilities and Subset Size

We evaluate the effectiveness of the various methods in terms of their probabilities of screening out inferior systems and retaining the best system, as well as the distributions of their subset sizes. We fix a problem instance with $k = 20$ systems and $\mu_i = -(1/4)(i-1)^{5/4}$ for $i = 1, \ldots, 20$, so that System 1 is the unique best (with $\mu_1 = 0$) and the expected performances become slightly more spread out as the systems get worse. A common sample size of $n_i = 5$ for $i = 1, \ldots, 20$ is assumed and the known variances are randomly generated as $\sigma_i^2 \sim \chi_{10}^2$ for $i = 1, \ldots, 20$, independent across systems and fixed across macroreplications.

We perform 10,000 macroreplications of each subset-selection method: generating sample data, computing indices or posterior probabilities, and forming subsets. On a given macroreplication, common random numbers are used across all methods, i.e., they form subsets based on the same sample data. The

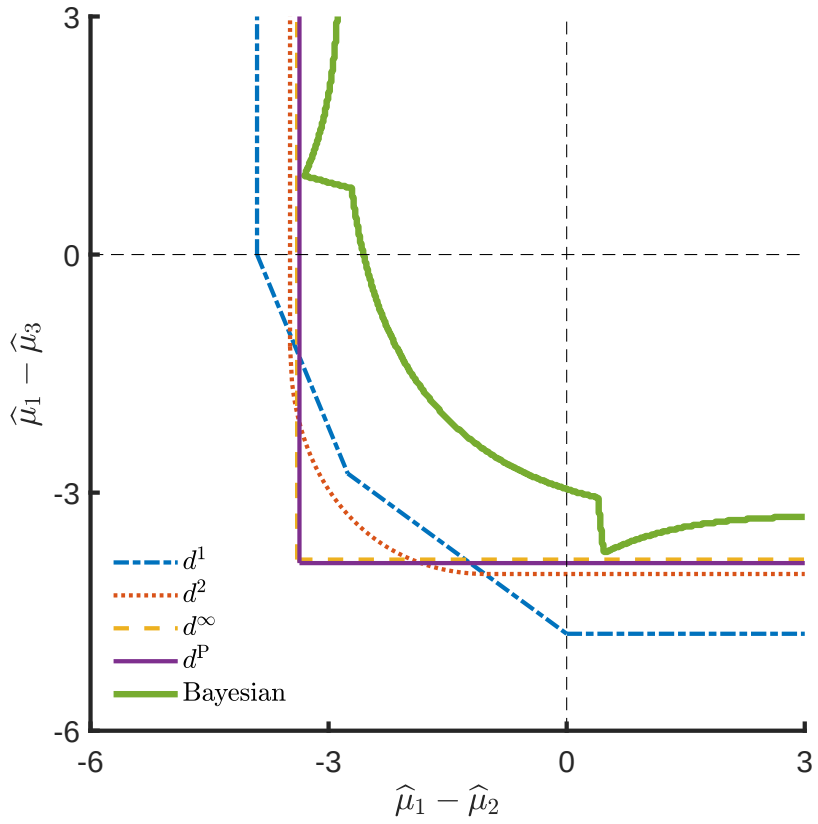

Figure 1: Boundaries of acceptance regions for including System 1 in the subset for the four standardized discrepancies and Bayesian subset selection for $k = 3$ systems and $1 - \alpha = 0.95$.

discrepancy-based methods are designed to deliver PCS guarantee (1) for $1 - \alpha = 0.95$, while the Bayesian subset-selection method likewise uses $1 - \alpha = 0.95$ as a target posterior PCS for its subset. The tightest cutoffs are estimated via Monte Carlo using 5000 replications; see Section 6 for full details.

Figure 2 shows the estimated probability that each system is included in the subset—estimates are each accurate to within ± 0.01 with 95% confidence. The reported inclusion probabilities are noticeably not monotone in the true means due to the random (therefore unordered) variances associated with the systems. The inclusion probabilities for the four standardized discrepancies tend to closely track across all systems, with an apparent ordering, from high to low, of d^1, d^2, d^P, and d^∞. The Bayesian subset-selection procedure, which we reiterate is not designed to deliver the frequentist PCS guarantee (1), has considerably smaller inclusion probabilities, especially for systems that are not far from the best. For each of the discrepancy-based methods, the probability of retaining the best system is at or above 99% (see Table 1), reflecting the inherent conservativeness of the PCS guarantee. In contrast, the Bayesian subset-selection procedure retains the best system on about 92% of the macroreplications, somewhat below the nominal coverage sought by the frequentist procedures.

Figure 3 shows the empirical cumulative distribution function (ecdf) of the subset size, $|S|$, for each method. Again, the estimated probabilities are accurate to within ± 0.01 with 95% confidence. The ecdfs of the d^∞ and d^P standardized discrepancies are virtually indistinguishable, with those of d^1 and d^2 being noticeably to the right. This indicates that the latter two standardized discrepancies are less effective at

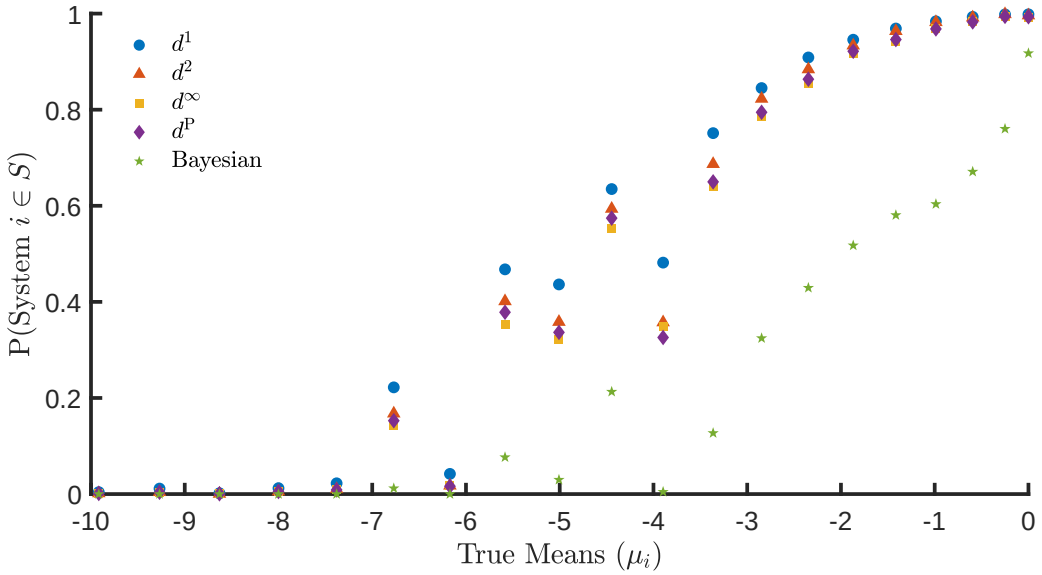

Figure 2: Inclusion probabilities for the four standardized discrepancies and Bayesian subset selection.

Table 1: Coverage of Best System and Average Subset Size

	d^1	d^2	d^∞	d^{P}	Bayesian
$\mathbb{P}(1 \in S)$	0.999 ± 0.001	0.996 ± 0.001	0.993 ± 0.002	0.993 ± 0.002	0.917 ± 0.005
$\mathbb{E}[\lvert S \rvert]$	10.73 ± 0.04	10.18 ± 0.03	9.84 ± 0.04	9.92 ± 0.4	5.27 ± 0.03

screening out inferior systems on this particular problem instance. Here too, we see that the Bayesian subset-selection method results in appreciably smaller subsets, as evidenced by the leftward shift in its ecdf relative to the others. In terms of average subset sizes, those of the discrepancy-based methods are about 10–11 systems while that of the Bayesian subset-selection procedure is roughly half of that, as reported in Table 1.

6 OPTIMAL CUTOFFS

We now turn to the problem of identifying a cutoff D_i satisfying Condition (3). Plumlee and Nelson (2018) derive a uniform cutoff, i.e., one to be used for all systems, by using the fact that if System i is optimal, then $\mu \in \mathsf{M}_i$. By setting $m = \mu$ in Definition (2), one obtains a random variable that upper bounds $D_i(\widehat{\mu})$ with probability 1 and whose distribution is independent of the unknown μ. This distribution also does not depend on the sampling variances for the standardized discrepancies we have considered, e.g., for the d^2 standardized discrepancy,

$$D_i^2(\widehat{\mu}) = \min_{m \in \mathsf{M}_i} \sum_{j=1}^{k} \frac{n_j}{\sigma_j^2}(\widehat{\mu}_j - m_j)^2 \leq \sum_{j=1}^{k} \frac{n_j}{\sigma_j^2}(\widehat{\mu}_j - \mu_j)^2 \stackrel{d}{=} \chi_k^2, \tag{9}$$

where χ_k^2 is a chi-squared random variable with k degrees of freedom and $\stackrel{d}{=}$ denotes equality in distribution. Thus a uniform cutoff of $\mathsf{D}_i^2 = \chi^2_{1-\alpha,k}$, the $1-\alpha$ quantile of a χ_k^2 random variable, ensures the PCS guarantee. Analogous uniform cutoffs can be derived for the d^1, d^∞, and d^{P} standardized discrepancies. When the bounding random variable does not follow a tractable parametric distribution, simulation can still be used to estimate the quantile, provided $D_i(\widehat{\mu})$ is cheap to compute.

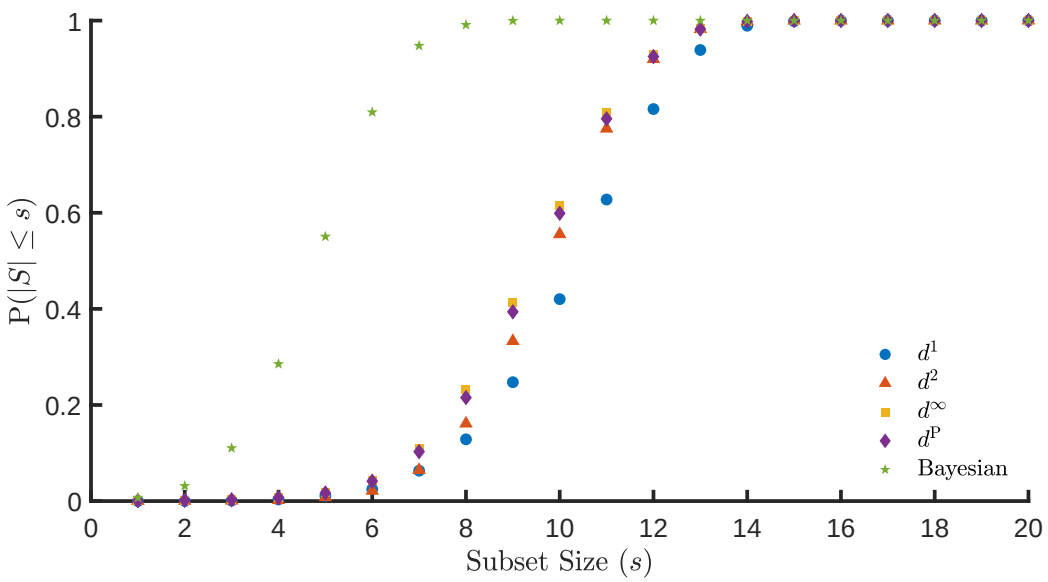

Figure 3: Empirical cumulative distribution function of the subset size for the four standardized discrepancies and Bayesian subset selection.

These uniform cutoffs tend to be overly conservative, with simulation experiments suggesting that for $1-\alpha = 0.95$, $\mathbb{P}([k] \in S) \geq 99.5\%$ for as few as $k = 10$ systems, with the overcoverage growing more extreme as the number of systems increases. Instead of a uniform cutoff, we investigate tighter cutoffs for each system that incorporate knowledge of the sampling variances. From Condition (3), we see that the smallest value D_i can take is the maximum $1-\alpha$ quantile of $D_i(\widehat{\mu})$ over all $\mu \in \mathsf{M}_i$. Although this would seem to be a difficult quantity to determine, Proposition 3 states that, under certain assumptions, one need only compute the $1-\alpha$ quantile of $D_i(\widehat{\mu})$ for $\mu = \mathbf{0}_k$.

Proposition 3 Let $\rho(i,\mu)$ denote the $1-\alpha$ quantile of $D_i(\widehat{\mu})$ given μ. If Assumption 1 and Conditions (C1) and (C2) hold, then $\max_{\mu\in\mathsf{M}_i} \rho(i,\mu) = \rho(i,\mathbf{0}_k)$.

Proof of Proposition 3. For ease of presentation, we temporarily let $D_i(\widehat{\mu},\mu)$ denote the random variable $D_i(\widehat{\mu})$ given the problem instance μ. Fix an arbitrary problem instance $\mu \in \mathsf{M}_i$ and express the sample means as $\widehat{\mu}_j = \mu_j + (\sigma_j/\sqrt{n_j})Z_j$ for $j = 1,\ldots,k$ where $Z_1,\ldots,Z_k$ are independent and identically distributed (i.i.d.) standard normal random variables. By Condition (C2), $D_i(\widehat{\mu},\mu)$ is monotone increasing in $\widehat{\mu}_j$ for all $j \neq i$. Therefore $D_i(\widehat{\mu},\mu)$ is first-order stochastically dominated by $D_i(\widehat{\mu},\mu_i\mathbf{1}_k)$. Consequently, $\rho(i,\mu) \leq \rho(i,\mu_i\mathbf{1}_k)$.

By Condition (C1), $d(m,\widehat{\mu}) = d(m-\mu_i\mathbf{1}_k, \widehat{\mu}-\mu_i\mathbf{1}_k)$. It follows that

$$D_i(\widehat{\mu},\mu_i\mathbf{1}_k) = \min_{m\in\mathsf{M}_i} d(m,\widehat{\mu}) = \min_{m\in\mathsf{M}_i} d(m-\mu_i\mathbf{1}_k, \widehat{\mu}-\mu_i\mathbf{1}_k) = \min_{m'\in\mathsf{M}_i} d(m',\widehat{\mu}-\mu_i\mathbf{1}_k),$$

where the last equality holds because for any $m \in \mathsf{M}_i$, $m-\mu_i\mathbf{1}_k \in \mathsf{M}_i$, i.e., subtracting μ_i from all of the values does not change the fact that the ith component is the largest. Then since, under problem instance $\mu_i\mathbf{1}_k$, $\widehat{\mu}_j = \mu_i + (\sigma_j/\sqrt{n_j})Z_j$ for all $j = 1,\ldots,k$, we have that $\widehat{\mu}-\mu_i\mathbf{1}_k = ((\sigma_1/\sqrt{n_1})Z_1,\ldots,(\sigma_k/\sqrt{n_k})Z_k)$. Therefore $D_i(\widehat{\mu},\mu_i\mathbf{1}_k) \stackrel{d}{=} D_i(\widehat{\mu},\mathbf{0}_k)$, implying that $\rho(i,\mu_i\mathbf{1}_k) = \rho(i,\mathbf{0}_k)$. Since the problem instance μ was arbitrary, $\max_{\mu\in\mathsf{M}_i} \rho(i,\mu) = \rho(i,\mathbf{0}_k)$. □

Proposition 3 implies that the tightest cutoffs for all solutions can be estimated simultaneously by running the following procedure:

1. Generate i.i.d. realizations of $\widehat{\mu}$ where $\widehat{\mu}_i \sim N(0,\sigma_i^2/n_i)$ for $i = 1,\ldots,k$, independent.

2. For each $\widehat{\mu}$, compute $D_i(\widehat{\mu})$ for all $i = 1, \ldots, k$.
3. For $i = 1, \ldots, k$, set D_i to be the empirical $1-\alpha$ quantile of the $D_i(\widehat{\mu})$ terms.

When the variances are unknown, uniform cutoffs can be derived in a similar fashion to Inequality (9). Obtaining tighter cutoffs, however, remains a challenging problem. The result of Proposition 3 that $\mu = \mathbf{0}_k$ represents a worst-case configuration of expected performances still holds, but μ no longer completely characterizes the problem instance. A worst-case configuration of sampling variances must also be determined if Step 1 of the estimation algorithm above is to be carried out. An inexact, yet practical, resolution is to nevertheless run the algorithm having plugged in the observed sample variances for the unknown quantities. We leave the task of exploring other methods for deriving valid, tighter cutoffs for future research.

7 CONCLUSION

Subset selection is just one approach to simulation optimization; for a general reference on simulation optimization, see Fu (2015). It is a workhorse screening method for output analysis due to its ability to accommodate unequal sample sizes and deliver a fixed-confidence guarantee without needing an indifference-zone specification. We present a general discrepancy-based framework for constructing subsets that deliver the frequentist PCS guarantee. This new paradigm offers insightful connections to existing subset-selection methods through different choices of standardized discrepancies and cutoffs. We also analyze properties of specific standardized discrepancies and derive simple expressions for the associated indices of systems. Experimental results suggest that standardized discrepancies motivated by the ℓ_∞ distance function yield smaller subset sizes for problem instances in which the expected performances are spread out.

The proposed framework generalizes several well-known subset-selection methods, yet others, e.g., Bayesian subset selection, cannot be easily recast in terms of discrepancies and cutoffs. Although the discrepancy-based methods do not naturally lend themselves to fully sequential procedures, splitting α across stages may offer an acceptable compromise. Future work in this area entails extensions to handle common random numbers and unknown variances, including finding tighter cutoffs.

ACKNOWLEDGMENTS

We thank Henry Lam for helpful conversations about the tightest-cutoff problem. This work was supported by the National Science Foundation Grant Number CMMI–1634982.

REFERENCES

Alam, K., and M. H. Rizvi. 1966. "Selection from Multivariate Normal Populations". *Annals of the Institute of Statistical Mathematics* 18(1):307–318.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. 1972. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. New York: John Wiley & Sons, Inc.

Boesel, J., B. L. Nelson, and S.-H. Kim. 2003. "Using Ranking and Selection to 'Clean up' after Simulation Optimization". *Operations Research* 51(5):814–825.

Boyd, S., and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge: Cambridge University Press.

DeGroot, M. H. 2004. *Optimal Statistical Decisions*. New York: John Wiley & Sons, Inc.

Eckman, D. J. 2019. *Reconsidering Ranking-and-Selection Guarantees*. Ph. D. thesis, Cornell University, Ithaca, New York.

Fu, M. 2015. *Handbook of Simulation Optimization*. New York: Springer.

Gupta, S. S. 1965. "On Some Multiple Decision (Selection and Ranking) Rules". *Technometrics* 7(2):225–245.

Hamilton, C., T. L. Bratcher, and J. D. Stamey. 2008. "Bayesian Subset Selection Approach to Ranking Normal Means". *Journal of Applied Statistics* 35(8):847–851.

Kim, S.-H., and B. L. Nelson. 2006. "Selecting the Best System". In *Simulation*, edited by S. G. Henderson and B. L. Nelson, Volume 13 of *Handbooks in Operations Research and Management Science*, Chapter 17, 501–534. Elsevier.

Miescke, K.-J. 1979. "Bayesian Subset Selection for Additive and Linear Loss Functions". *Communications in Statistics–Theory and Methods* 8(12):1205–1226.

Nelson, B. L., J. Swann, D. Goldsman, and W. Song. 2001. "Simple Procedures for Selecting the Best Simulated System When the Number of Alternatives Is Large". *Operations Research* 49(6):950–963.

Pei, L., B. L. Nelson, and S. Hunter. 2018. "A New Framework for Parallel Ranking & Selection Using an Adaptive Standard". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 2201–2212. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Plumlee, M., and B. L. Nelson. 2018. "Plausible Optima". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1981–1992. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Robertson, T., F. T. Wright, and R. L. Dykstra. 1988. *Order Restricted Statistical Inference*. New York: John Wiley & Sons, Inc.

Silvapulle, M. J., and P. K. Sen. 2005. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Hoboken, New Jersey: John Wiley & Sons, Inc.

AUTHOR BIOGRAPHIES

DAVID J. ECKMAN is a postdoctoral research fellow in the Department of Industrial Engineering and Management Sciences at Northwestern University. His research interests deal with optimization and output analysis for stochastic simulation models. His e-mail address is david.eckman@northwestern.edu.

MATTHEW PLUMLEE is an Assistant Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He primarily researches uncertainty quantification methods for computational models of systems. His e-mail address is mplumlee@northwestern.edu.

BARRY L. NELSON is the Walter P. Murphy Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is a Fellow of INFORMS and IISE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of *Foundations and Methods of Stochastic Simulation: A First Course*, from Springer. His e-mail address is nelsonb@northwestern.edu.