Selecting a Selection Procedure

Jürgen Branke* Stephen E. Chick† Christian Schmidt*

Abstract

Selection procedures are used in a variety of applications to select the best of a finite set of alternatives. 'Best' is defined with respect to the largest mean, but the mean is inferred with statistical sampling, as in simulation optimization. There are a wide variety of procedures, which begs the question of which selection procedure to select. The main contribution of this paper is to identify, through extensive experimentation, the most effective selection procedures when samples are independent and normally distributed. We also (a) summarize the main structural approaches to deriving selection procedures, (b) formalize new sampling allocations and stopping rules, (c) identify strengths and weaknesses of the procedures, (d) identify some theoretical links between them, (e) and present an innovative empirical test bed with the most extensive numerical comparison of selection procedures to date. The most efficient and easiest to control procedures allocate samples with a Bayesian model for uncertainty about the means, and use new adaptive stopping rules proposed here.

Selection procedures are intended to select the best of a finite set of alternatives, where best is determined with respect to the largest mean, but the mean must be inferred via statistical sampling (Bechhofer et al. 1995). Selection procedures can inform managers how to select the best of a small set of alternative actions whose effects are evaluated with simulation (Nelson and Goldsman 2001), and have been implemented in commercial simulation products. Selection procedures have also attracted interest in combination with tools like multiple attribute utility theory (Butler et al. 2001), evolutionary algorithms (Branke and Schmidt 2004), and discrete optimization via simulation (Boesel et al. 2003).

Three main approaches to solving the selection problem are distinguished by their assumptions about how the evidence for correct selection is described and sampling allocations are made: the indifference zone (IZ, Kim and Nelson 2006), the expected value of information procedure (VIP, Chick and Inoue 2001a), and the optimal computing budget allocation (OCBA, Chen 1996) approaches. IZ procedures typically allocate samples in order to provide a guaranteed lower bound for the frequentist probability of correct selection (PCS), with respect to the sampling distribution, for selection problems in a specific class (e.g., the mean of the best is at least a prespecified amount better than each alternative). The VIP approach describes the evidence for correct selection with Bayesian posterior distributions, and allocates further samples using decision-theory tools to maximize the expected value of information in those samples. The OCBA is a heuristic that uses a normal distribution approximation for the Bayesian posterior distribution of the unknown mean performance of each alternative in order to sequentially allocate further samples. Each approach stipulates a number of different sampling assumptions, approximations, stopping rules and parameters that combine to define a procedure. With so many variations, the question of which selection procedure to select arises. The question is important because the demands that are being placed upon simulation optimization algorithms are increasing. The answer may also provide new insights about differences in the empirical performance of three distinct approaches to statistical decision making (classical, or frequentist, statistics; Bayesian decision theory; and heuristic models about the probability of correct selection).

*Institute AIFB, University of Karlsruhe, Germany {branke, csc}@aifb.uni-karlsruhe.de

†INSEAD, Technology and Operations Management Area, stephen.chick@insead.edu

A thorough comparison of these three approaches has not previously been done. Initial work shows that special cases of the VIP outperform specific IZ and OCBA procedures (in a comparison of two-stage procedures), and specific sequential VIP and OCBA procedures are more efficient than two-stage procedures (Inoue et al. 1999). The $\mathcal{KN}$ family of procedures is effective among IZ procedures (Kim and Nelson 2006). No paper has studied more than a limited set of procedures with respect to a moderate experimental test bed.

This paper addresses the unmet need for an extensive comparison of IZ, VIP and OCBA procedures. §1 summarizes the main approaches to selection procedures, derives new variants and formalizes new stopping rules for the VIP and OCBA procedures. Each procedure makes approximations, and none provides an optimal solution, so it is important to understand the strengths and weaknesses of each approach. §2 describes new measurements to evaluate each with respect to:

- Efficiency: The mean evidence for correct selection as a function of the mean number of samples.
- Controllability: The ease of setting a procedure's parameters to achieve a targeted evidence level.
- Robustness: The dependency of a procedure's effectiveness on the underlying problem characteristics.
- Sensitivity: The effect of the parameters on the mean number of samples needed.

Some practitioners desire a (statistically conservative) lower bound for the targeted evidence level, such as a frequentist $\mathrm{PCS}_{\mathrm{IZ}}$ guarantee, but this may lead to excessive sampling. Together, efficiency and controllability indicate how close to the desired evidence level a procedure gets while avoiding excess sampling.

The procedures are compared empirically on a large variety of selection problems described in §3. The test bed is unique not only because of its size, but also by its inclusion of randomized problem instances, in addition to structured problem instances that are usually studied, but that are unlikely to be found in practice.

The focus is on applications where the samples are jointly independent and normally distributed with unknown and potentially different variances, or nearly so as is the case in stochastic simulation with batching (Law and Kelton 2000). Branke et al. (2005) presented a subset of preliminary empirical results, and assessed additional stopping rules that were somewhat less efficient than those considered below.

§4 empirically compares the different selection procedures on a variety of test problems. The results show that a leading IZ procedure, called $\mathcal{KN}{++}$ (described below), is more efficient than the original VIP and OCBA procedures, but is statistically conservative which may result in excessive sampling. In combination with the new stopping rules, the VIP and OCBA procedures are most efficient. They also tend to be more controllable and robust in the experiments below. §5 recommends those procedures, and discusses key issues for selecting a selection procedure. Appendices in the Online Companion generalize an OCBA procedure, give structural results that suggest why certain VIP and OCBA procedures perform similarly, describe the implementation, and display and interpret additional numerical results.

1 The Procedures

We first formalize the problem, summarize assumptions and establish notation. §1.1 describes measures of the evidence of correct selection and, based thereon, introduces new stopping rules that improve efficiency. §1.2-1.4 describe existing and new procedures from the IZ, VIP and OCBA approaches.

The best of k simulated systems is to be identified, where 'best' means the largest output mean. Analogous results hold if smallest is best. Let X_{ij} be a random variable whose realization x_{ij} is the output of the jth simulation replication of system i, for $i = 1, \ldots, k$ and $j = 1, 2, \ldots$. Let w_i and σ_i^2 be the unknown mean and variance of simulated system i, and let $w_{[1]} \leq w_{[2]} \leq \ldots \leq w_{[k]}$ be the ordered means. In practice, the ordering $[\cdot]$ is unknown, and the best system, system $[k]$, is to be identified with simulation. The procedures considered below are derived from the assumption that simulation output is independent and normally distributed, *conditional* on w_i and σ_i^2, for $i = 1, \ldots, k$.

$$\{X_{ij} : j = 1, 2, \ldots\} \overset{iid}{\sim} \texttt{Normal}\left(w_i, \sigma_i^2\right)$$

Although the normality assumption is not always valid, it is often possible to batch a number of outputs so that normality is approximately satisfied. Vectors are written in boldface, such as $\mathbf{w} = (w_1, \ldots, w_k)$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)$. A problem instance (*configuration*) is denoted by $\boldsymbol{\chi} = (\mathbf{w}, \boldsymbol{\sigma}^2)$.

Let n_i be the number of replications for system i run so far. Let $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ be the sample mean and $\hat{\sigma}_i^2 = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2/(n_i - 1)$ be the sample variance. Let $\bar{x}_{(1)} \leq \bar{x}_{(2)} \leq \ldots \leq \bar{x}_{(k)}$ be the ordering of the sample means based on all replications seen so far. Equality occurs with probability 0 in contexts of interest here. The quantities $n_i, \bar{x}_i, \hat{\sigma}_i^2$ and (i) are updated as more replications are observed.

Each selection procedure generates estimates $\hat{w}_i$ of w_i, for $i = 1, \ldots, k$. For the procedures studied here, $\hat{w}_i = \bar{x}_i$, and a correct selection occurs when the selected system, system $\mathfrak{D}$, is the best system, $[k]$. Usually $\mathfrak{D} = (k)$ is selected as best.

If T_ν is a random variable with standard t distribution with ν degrees of freedom, we denote (as do Bernardo and Smith 1994) the distribution of $\mu + \frac{1}{\sqrt{\kappa}} T_\nu$ by $\texttt{St}\,(\mu, \kappa, \nu)$. If $\nu > 2$ the variance is $\kappa^{-1}\nu/(\nu - 2)$. If $\kappa = \infty$ or 1/0, then $\texttt{St}\,(\mu, \kappa, \nu)$ denotes a point mass at μ. Denote the cumulative distribution function (cdf) of the standard t distribution ($\mu = 0, \kappa = 1$) by $\Phi_\nu()$ and probability density function (pdf) by $\phi_\nu()$.

1.1 Evidence for Correct Selection

This section provides a unified framework for describing both frequentist and Bayesian measures of selection procedure effectiveness and the evidence of correct selection. They are required to derive and compare the procedures below. They are also used within the Bayesian procedures (VIP and OCBA) to decide when the evidence of correct selection is sufficient to stop sampling.

The measures are defined in terms of loss functions. The zero-one loss function, $\mathcal{L}_{0-1}(\mathfrak{D}, \mathbf{w}) = \mathbb{1}\left\{w_{\mathfrak{D}} \neq w_{[k]}\right\}$, equals 1 if the best system is not correctly selected, and is 0 otherwise. The opportunity cost $\mathcal{L}_{oc}(\mathfrak{D}, \mathbf{w}) = w_{[k]} - w_{\mathfrak{D}}$ is 0 if the best system is correctly selected, and is otherwise the difference between the best and selected system. The opportunity cost makes more sense in business applications.

The IZ procedures take a frequentist perspective. The frequentist probability of correct selection ($\mathrm{PCS}_{\mathrm{IZ}}$) is the probability that the mean of the system selected as best, system $\mathfrak{D}$ equals the mean of the system with the highest mean, system $[k]$, conditional on the problem instance (this allows for ties). The probability is with respect to the simulation output X_{ij} generated by the procedure (the realizations x_{ij} determine $\mathfrak{D}$).

$$\mathrm{PCS}_{\mathrm{IZ}}(\boldsymbol{\chi}) \overset{\text{def}}{=} 1 - \mathrm{E}\left[\mathcal{L}_{0-1}(\mathfrak{D}, \mathbf{w}) \,|\, \boldsymbol{\chi}\right] = \Pr\left(w_{\mathfrak{D}} = w_{[k]} \,|\, \boldsymbol{\chi}\right)$$

Indifference zone procedures attempt to guarantee a lower bound on $\mathrm{PCS}_{\mathrm{IZ}}$, subject to the indifference-zone constraint that the best system is at least $\delta^* > 0$ better than the others,

$$\mathrm{PCS}_{\mathrm{IZ}}(\boldsymbol{\chi}) \geq 1 - \alpha^*, \text{for all } \boldsymbol{\chi} = (\mathbf{w}, \boldsymbol{\sigma}^2) \text{ such that } w_{[k]} \geq w_{[k-1]} + \delta^*. \tag{1}$$

A selected system within δ^* of the best is called *good*. Some IZ procedures satisfy frequentist probability of good selection guarantees, $\mathrm{PGS}_{\mathrm{IZ},\delta^*}(\boldsymbol{\chi}) \overset{\text{def}}{=} \Pr\left(w_{\mathfrak{D}} > w_{[k]} - \delta^* \,|\, \boldsymbol{\chi}\right) \geq 1 - \alpha^*$, for *all* configurations (Nelson and Banerjee 2001). Let $\mathrm{PICS}_{\mathrm{IZ}} = 1 - \mathrm{PCS}_{\mathrm{IZ}}$ and $\mathrm{PBS}_{\mathrm{IZ},\delta^*} = 1 - \mathrm{PGS}_{\mathrm{IZ},\delta^*}$ denote the probability of *incorrect* and *bad* selections.

An alternative to a PCS guarantee for the evidence of correct selection is a guaranteed upper bound on the expected opportunity cost (EOC) of a potentially incorrect selection. The frequentist EOC (Chick and Wu 2005) is also defined with respect to the sampling distribution,

$$\mathrm{EOC}_{\mathrm{IZ}}(\boldsymbol{\chi}) \overset{\text{def}}{=} \mathrm{E}\left[\mathcal{L}_{oc}(\mathfrak{D}, \mathbf{w}) \,|\, \boldsymbol{\chi}\right] = \mathrm{E}\left[w_{[k]} - w_{\mathfrak{D}} \,|\, \boldsymbol{\chi}\right].$$

Bayesian procedures assume that parameters whose values are unknown are random variables (such as the unknown means $\mathbf{W}$), and use the posterior distributions of the unknown parameters to measure the quality of a selection. Given the data $\mathcal{E}$ seen so far, two measures of selection quality are

$$\begin{aligned} \mathrm{PCS}_{\mathrm{Bayes}} &\overset{\text{def}}{=} 1 - \mathrm{E}\left[\mathcal{L}_{0-1}(\mathfrak{D}, \mathbf{W}) \,|\, \mathcal{E}\right] = \Pr\left(W_{\mathfrak{D}} \geq \max_{i \neq \mathfrak{D}} W_i \,|\, \mathcal{E}\right) \\ \mathrm{EOC}_{\mathrm{Bayes}} &\overset{\text{def}}{=} \mathrm{E}\left[\mathcal{L}_{oc}(\mathfrak{D}, \mathbf{W}) \,|\, \mathcal{E}\right] = \mathrm{E}\left[\max_{i=1,2,\ldots,k} W_i - W_{\mathfrak{D}} \,|\, \mathcal{E}\right], \end{aligned} \tag{2}$$

the expectation taken over both $\mathfrak{D}$ (which is determined by the random X_{ij}) and the posterior distribution of $\mathbf{W}$, given $\mathcal{E}$. Assuming a noninformative prior distribution for the unknown mean and variance, the posterior marginal distribution for the unknown mean W_i, given $n_i > 2$ samples, is $\mathtt{St}\left(\bar{x}_i, n_i/\hat{\sigma}_i^2, \nu_i\right)$, where $\nu_i = n_i - 1$ (de Groot 1970). Each Bayesian procedure below selects the system with the best sample mean after all sampling is done, $\mathfrak{D} = (k)$.

Approximations in the form of bounds on the above losses are useful to derive sampling allocations and to define stopping rules. *Slepian's inequality* (e.g., see Kim and Nelson 2006) implies that the posterior evidence that system (k) is best satisfies

$$\mathrm{PCS}_{\mathrm{Bayes}} \geq \prod_{j:(j) \neq (k)} \Pr\left(W_{(k)} > W_{(j)} \,|\, \mathcal{E}\right). \tag{3}$$

The right hand side of Inequality (3) is approximately

$$\mathrm{PCS}_{\mathrm{Slep}} = \prod_{j:(j) \neq (k)} \Phi_{\nu_{(j)(k)}}(d^*_{jk}),$$

where d^*_{jk} is the normalized distance for systems (j) and (k), and $\nu_{(j)(k)}$ comes from *Welch's approximation* for the difference $W_{(k)} - W_{(j)}$ of two shifted and scaled t random variables (Law and Kelton 2000, p. 559):

$$\begin{aligned} d^*_{jk} &= d_{(j)(k)} \lambda_{jk}^{1/2} \text{ with } d_{(j)(k)} = \bar{x}_{(k)} - \bar{x}_{(j)} \text{ and } \lambda_{jk}^{-1} = \frac{\hat{\sigma}^2_{(j)}}{n_{(j)}} + \frac{\hat{\sigma}^2_{(k)}}{n_{(k)}}, \\ \nu_{(j)(k)} &= \frac{[\hat{\sigma}^2_{(j)}/n_{(j)} + \hat{\sigma}^2_{(k)}/n_{(k)}]^2}{[\hat{\sigma}^2_{(j)}/n_{(j)}]^2/(n_{(j)} - 1) + [\hat{\sigma}^2_{(k)}/n_{(k)}]^2/(n_{(k)} - 1)}. \end{aligned} \tag{4}$$

We found that the Welch approximation outperformed another approximation in earlier comparisons of selection procedures (Branke et al. 2005). The Bayesian posterior probability of a good selection, where the selected system is within δ^* of the best, can be approximated in a similar manner by

$$\mathrm{PGS}_{\mathrm{Slep},\delta^*} = \prod_{j:(j)\neq(k)} \Phi_{\nu_{(j)(k)}}(\lambda_{jk}^{1/2}(\delta^* + d_{(j)(k)})). \tag{5}$$

The term $\mathrm{EOC}_{\mathrm{Bayes}}$ may be expensive to compute if $k > 2$. Summing the losses from $(k-1)$ pairwise comparisons between the current best and each other system gives an easily computed upper bound (Chick and Inoue 2001a, 2002). Let $f_{(j)(k)}(\cdot)$ be the posterior pdf for the difference $W_{(j)} - W_{(k)}$, given $\mathcal{E}$. This is, approximately, a $\mathtt{St}\left(-d_{(j)(k)}, \lambda_{jk}, \nu_{(j)(k)}\right)$ distribution. We denote the standardized EOC function by

$$\Psi_\nu[s] \stackrel{\mathrm{def}}{=} \int_{u=s}^{\infty} (u-s)\phi_\nu(u)du = \frac{\nu + s^2}{\nu - 1}\phi_\nu(s) - s\Phi_\nu(-s). \tag{6}$$

Chick and Inoue's upper bound for $\mathrm{EOC}_{\mathrm{Bayes}}$ is easy to approximate, using Bonferroni's inequality, by $\mathrm{EOC}_{\mathrm{Bonf}}$, where

$$\begin{aligned}\mathrm{EOC}_{\mathrm{Bayes}} &\leq \sum_{j:(j)\neq(k)} \int_{w=0}^{\infty} w\, f_{(j)(k)}(w)\, dw \\ &\approx \sum_{j:(j)\neq(k)} \lambda_{jk}^{-1/2} \Psi_{\nu_{(j)(k)}}\left[d_{jk}^*\right] \stackrel{\mathrm{def}}{=} \mathrm{EOC}_{\mathrm{Bonf}} \end{aligned} \tag{7}$$

The VIP and OCBA procedures defined below can use the values of $\mathrm{EOC}_{\mathrm{Bonf}}$ and $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ to decide when to stop sampling. In particular, the following **stopping rules** are used:

1. Sequential ($\mathcal{S}$): Repeat sampling while $\sum_{i=1}^{k} n_i < B$ for some specified total budget B.

2. Probability of good selection ($\mathrm{PGS}_{\mathrm{Slep},\delta^*}$): Repeat while $\mathrm{PGS}_{\mathrm{Slep},\delta^*} < 1 - \alpha^*$ for a specified probability target $1 - \alpha^*$ and given $\delta^* \geq 0$.

3. Expected opportunity cost ($\mathrm{EOC}_{\mathrm{Bonf}}$): Repeat while $\mathrm{EOC}_{\mathrm{Bonf}} > \beta^*$, for a specified EOC target β^*.

Prior work for sequential VIP and OCBA procedures used the $\mathcal{S}$ stopping rule. The other stopping rules provide the flexibility to stop earlier if the evidence for correct selection is sufficiently high, and allow for additional sampling when the evidence is not sufficiently high. The IZ requires $\delta^* > 0$. The VIP and OCBA permit $\delta^* = 0$ to obtain a pure PCS-based stopping condition. We use $\mathrm{PCS}_{\mathrm{Slep}}$ to denote $\mathrm{PGS}_{\mathrm{Slep},0}$.

1.2 Indifference Zone (IZ)

The IZ approach (Bechhofer et al. 1995; Kim and Nelson 2006) seeks to guarantee $\mathrm{PCS}_{\mathrm{IZ}} \geq 1 - \alpha^* > 1/k$, whenever the best system is at least δ^* better than the other systems. The indifference-zone parameter δ^* is typically elicited as the smallest difference in mean performance that is significant to the decision-maker.

Early IZ procedures were statistically conservative in the sense of excess sampling unless unfavorable configurations of the means were found. The $\mathcal{KN}$ family of procedures, which might be considered state of the art for the IZ approach, improves sampling efficiency over a broad set of configurations (Kim and Nelson

2001). The original $\mathcal{KN}$ procedure provides a PCS guarantee, and estimates the variance of the output of each system with the sample variances from a first stage of sampling alone. Procedure $\mathcal{KN}$++ (Goldsman et al. 2002) updates the sample variance as more samples are observed, but only provides an asymptotic PCS guarantee as $\delta^* \to 0$. That asymptotic guarantee is also valid with nonnormal samples. We found that more frequent updates of the sample variance increases efficiency (see Appendix A.1).

Some $\mathcal{KN}$ procedures, including $\mathcal{KN}$++, can handle the more general case of correlated simulation output. Here we specialize Procedure $\mathcal{KN}$++ for *independent* output. The procedure screens out some systems as runs are observed, and each noneliminated system is simulated the same number of times. We used $\xi = 1$ replication per stage per noneliminated system, and updated the sample variance in every stage.

Procedure $\mathcal{KN}$++ (independent samples)

1. Specify a confidence level $1 - \alpha^* > 1/k$, an indifference-zone parameter $\delta^* > 0$, a first-stage sample size $n_0 > 2$, and a number ξ of samples to run per noneliminated system per subsequent stage.
2. Initialize the set of noneliminated systems, $I \leftarrow \{1, \ldots, k\}$, set $n \leftarrow 0, \tau \leftarrow n_0$.
3. WHILE $|I| > 1$ DO another stage:
   (a) Observe τ additional samples from system i, for all $i \in I$. Set $n \leftarrow n + \tau$. Set $\tau \leftarrow \xi$.
   (b) Update: Set $\eta \leftarrow \frac{1}{2}\left\{[2(1-(1-\alpha^*)^{1/(k-1)})]^{-2/(n-1)} - 1\right\}$ and $h^2 \leftarrow 2\eta(n-1)$. For all $i \in I$, update the sample statistics $\bar{x}_i$ and $\hat{\sigma}_i^2$.
   (c) Screen: For all $i, j \in I$ and $i > j$, set $d_{ij} \leftarrow \bar{x}_j - \bar{x}_i$ and $\epsilon_{ij} \leftarrow \max\left\{0, \frac{\delta^*}{2n}\left(\frac{h^2(\hat{\sigma}_i^2+\hat{\sigma}_j^2)}{\delta^{*2}} - n\right)\right\}$. If $d_{ij} > \epsilon_{ij}$ then $I \leftarrow I \backslash \{i\}$. If $d_{ij} < -\epsilon_{ij}$ then $I \leftarrow I \backslash \{j\}$.
4. Return remaining system, system $\mathfrak{D}$, as best.

1.3 Value of Information Procedure (VIP)

VIPs allocate samples to each alternative in order to maximize the expected value of information (EVI) of those samples. Some balance the cost of sampling with the EVI, and some maximize EVI subject to a sampling budget constraint (Chick and Inoue 2001a). Procedures 0-1($\mathcal{S}$) and $\mathcal{LL}(\mathcal{S})$ are sequential variations of those procedures that improve Bonferroni bounds for $\mathrm{PCS}_{\mathrm{Bayes}}$ (the expected 0-1 loss) and $\mathrm{EOC}_{\mathrm{Bayes}}$, respectively. An alternative name for EOC is linear loss – hence the name $\mathcal{LL}$. Those procedures allocate τ replications per stage until a total of B replications are run. The derivation of those procedures assumes that samples are normally distributed, but the general VIP framework can apply for nonnormal samples too.

This section recalls those procedures, and adapts them to permit the use of the stopping rules in §1.1.

Procedure 0-1.

1. Specify a first-stage sample size $n_0 > 2$, and a total number of samples $\tau > 0$ to allocate per subsequent stage. Specify stopping rule parameters.
2. Run independent replications $X_{i1}, \ldots, X_{in_0}$, and initialize the number of replications $n_i \leftarrow n_0$ run so far for each system, $i = 1, \ldots, k$.

3. Determine the sample statistics $\bar{x}_i$ and $\hat{\sigma}_i^2$, and the order statistics, so that $\bar{x}_{(1)} \leq \ldots \leq \bar{x}_{(k)}$.
4. WHILE stopping rule not satisfied DO another stage:
    (a) Initialize the set of systems considered for additional replications, $\mathcal{S} \leftarrow \{1, \ldots, k\}$.
    (b) For each (i) in $\mathcal{S}\backslash\{(k)\}$: If $(k) \in \mathcal{S}$ then set $\lambda_{ik}^{-1} \leftarrow \hat{\sigma}_{(i)}^2/n_{(i)} + \hat{\sigma}_{(k)}^2/n_{(k)}$, and set $\nu_{(i)(k)}$ with Welch's approximation. If $(k) \notin \mathcal{S}$ then set $\lambda_{ik} \leftarrow n_{(i)}/\hat{\sigma}_{(i)}^2$ and $\nu_{(i)(k)} \leftarrow n_{(i)} - 1$.
    (c) Tentatively allocate a total of τ replications to systems $(i) \in \mathcal{S}$ (set $\tau_{(j)} \leftarrow 0$ for $(j) \notin \mathcal{S}$):
    $$\tau_{(i)} \leftarrow \frac{(\tau + \sum_{j \in \mathcal{S}} n_j)(\hat{\sigma}_{(i)}^2 \gamma_{(i)})^{\frac{1}{2}}}{\sum_{j \in \mathcal{S}} (\hat{\sigma}_j^2 \gamma_j)^{\frac{1}{2}}} - n_{(i)}, \text{ where } \gamma_{(i)} \leftarrow \begin{cases} \lambda_{ik} d_{ik}^* \phi_{\nu_{(i)(k)}}(d_{ik}^*) & \text{for } (i) \neq (k) \\ \sum_{(j) \in \mathcal{S}\backslash\{(k)\}} \gamma_{(j)} & \text{for } (i) = (k) \end{cases}$$
    and the normalized distance d_{ik}^* is as in (4).
    (d) If any $\tau_i < 0$ then fix the nonnegativity constraint violation: remove (i) from $\mathcal{S}$ for each (i) such that $\tau_{(i)} \leq 0$, and go to Step 4b. Otherwise, round the τ_i so that $\sum_{i=1}^k \tau_i = \tau$ and go to Step 4e.
    (e) Run τ_i additional replications for system i, for $i = 1, \ldots, k$. Update sample statistics $n_i \leftarrow n_i + \tau_i$; $\bar{x}_i$; $\hat{\sigma}_i^2$, and the order statistics, so $\bar{x}_{(1)} \leq \ldots \leq \bar{x}_{(k)}$.
5. Select the system with the best estimated mean, $\mathfrak{D} = (k)$.

The formulas in Steps 4b-4c are derived from optimality conditions to improve a Bonferroni-like bound on the EVI for asymptotically large τ (Chick and Inoue 2001a). Depending on the stopping rule used, the resulting procedures are named 0-1($\mathcal{S}$), 0-1($\mathrm{PGS}_{\mathrm{Slep},\delta^*}$), 0-1($\mathrm{EOC}_{\mathrm{Bonf}}$), with the stopping rule in parentheses.

Procedure $\mathcal{LL}$ is a variant of 0-1 where sampling allocations seek to minimize $\mathrm{EOC}_{\mathrm{Bonf}}$.

Procedure $\mathcal{LL}$. Same as Procedure 0-1, except set $\gamma_{(i)}$ in Step 4c to

$$\gamma_{(i)} \leftarrow \begin{cases} \lambda_{ik}^{1/2} \frac{\nu_{(i)(k)} + (d_{ik}^*)^2}{\nu_{(i)(k)} - 1} \phi_{\nu_{(i)(k)}}(d_{ik}^*) & \text{for } (i) \neq (k) \\ \sum_{(j) \in \mathcal{S}\backslash\{(k)\}} \gamma_{(j)} & \text{for } (i) = (k) \end{cases} \tag{8}$$

1.4 OCBA Procedures

The OCBA is a class of procedures that was initially proposed by Chen (1996) and that has several variations (e.g. Chen et al. 2000; Chen et al. 2006). The variations involve different approximations for $\mathrm{PCS}_{\mathrm{Bayes}}$, and heuristics for how additional samples might improve the probability of correct selection. Here we specify the idea behind the OCBA and the variations used for this paper.

The OCBA assumes that if an additional τ replications are allocated for system i, but none are allocated for the other systems, then the standard error for the estimated mean of system i is scaled back accordingly. The usual OCBA assumes normal distributions to approximate the effect, but we use Student distributions,

$$\begin{aligned} \tilde{W}_i &\sim \mathtt{St}\left(\bar{x}_i, (n_i + \tau)/\hat{\sigma}_i^2, n_i - 1 + \tau\right) \\ \tilde{W}_j &\sim \mathtt{St}\left(\bar{x}_j, n_j/\hat{\sigma}_j^2, n_j - 1\right) \qquad \text{for } j \neq i, \end{aligned}$$

for consistency with a Bayesian assumption for the unknown σ_i^2. Chen et al. (2006) and Branke et al. (2005) found no notable difference in performance when comparing normal versus Student distributions for the $\tilde{W}_i$.

The effect of allocating an additional τ replications to system i, but no replications to the others, leads to an *estimated approximate probability of correct selection* (EAPCS) evaluated with respect to the distribution of $\tilde{\mathbf{W}} = (\tilde{W}_1, \ldots, \tilde{W}_k)$, and with $\tilde{W}_{(j)} - \tilde{W}_{(k)}$ approximated using Welch's approximation.

$$\begin{aligned} \mathrm{EAPCS}_i &= \prod_{j:(j)\neq(k)} \Pr\left(\tilde{W}_{(j)} < \tilde{W}_{(k)} \,|\, \mathcal{E}\right) \\ &\approx \prod_{j:(j)\neq(k)} (1 - \Phi_{\tilde{\nu}_{(j)(k)}}(\tilde{\lambda}_{jk}^{1/2} d_{(j)(k)})) \\ \tilde{\lambda}_{jk}^{-1} &= \frac{\hat{\sigma}_{(k)}^2}{n_{(k)} + \tau \mathbb{1}\left\{(k) = i\right\}} + \frac{\hat{\sigma}_{(j)}^2}{n_{(j)} + \tau \mathbb{1}\left\{(j) = i\right\}} \end{aligned} \tag{9}$$

where $\mathbb{1}\left\{\cdot\right\}$ is 1 if the argument is true, and 0 otherwise.

These approximations result in a sequential OCBA algorithm that greedily allocates samples to systems that most increase $\mathrm{EAPCS}_i - \mathrm{PCS}_{\mathrm{Slep}}$ at each stage. An innovation for the OCBA here is that sampling continues until a stopping rule from §1.1 is satisfied.

Procedure $\mathcal{OCBA}$.

1. Specify a first-stage sample size $n_0 > 2$, a number q of systems to simulate per stage, a sampling increment $\tau > 0$ to allocate per subsequent stage, and stopping rule parameters.
2. Run independent replications $X_{i1}, \ldots, X_{in_0}$, and initialize the number of replications $n_i \leftarrow n_0$ run so far for each system, $i = 1, \ldots, k$.
3. Determine the sample statistics $\bar{x}_i$ and $\hat{\sigma}_i^2$ and the sample mean ordering, so that $\bar{x}_{(1)} \leq \ldots \leq \bar{x}_{(k)}$.
4. WHILE stopping rule not satisfied DO another stage:
   (a) Compute EAPCS_i for $i = 1, \ldots, k$.
   (b) Set $\tau_i \leftarrow \tau/q$ for the q systems with largest $\mathrm{EAPCS}_i - \mathrm{PCS}_{\mathrm{Slep}}$, set $\tau_j \leftarrow 0$ for the others.
   (c) Run τ_i additional observations from system i.
   (d) For all i with $\tau_i > 0$, update $n_i \leftarrow n_i + \tau_i$, the sample statistics $\bar{x}_i$, $\hat{\sigma}_i^2$, and order statistics, so that $\bar{x}_{(1)} \leq \ldots \leq \bar{x}_{(k)}$.
5. Select the system with the best estimated mean, $\mathfrak{D} = (k)$.

He, Chick, and Chen (2006) proposed an OCBA variation, $\mathcal{OCBA}_{\mathrm{LL}}$, that accounts for the expected opportunity cost, and showed that the original $\mathcal{OCBA}$ procedure, the new $\mathcal{OCBA}_{\mathrm{LL}}$ and $\mathcal{LL}$ perform better than some other procedures in several empirical tests. By analogy with EAPCS_i above, set

$$\mathrm{EEOCS}_i = \sum_{j:(j)\neq(k)} \tilde{\lambda}_{jk}^{-1/2} \Psi_{\tilde{\nu}_{(j)(k)}} \left[\tilde{\lambda}_{jk}^{1/2} d_{(j)(k)}\right].$$

Procedure $\mathcal{OCBA}_{\mathrm{LL}}$ is a variation of $\mathcal{OCBA}$ that allocates replications to systems that maximize the improvement in expected opportunity cost (linear loss), $\mathrm{EOC}_{\mathrm{Bonf}} - \mathrm{EEOCS}_i$, in Step 4b.

Yet another OCBA heuristic incorporates the indifference zone parameter δ^* into the sampling allocation, not just the stopping rule. Let $\mathrm{EAPGS}_{i,\delta^*}$ generalize $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ by computing it with respect to the distribution of $\tilde{\mathbf{W}}$. **Procedure** $\mathcal{OCBA}_{\delta^*}$ allocates replications to systems that most improve an estimated probability of a good selection, $\mathrm{EAPGS}_{i,\delta^*} - \mathrm{PGS}_{\mathrm{Slep},\delta^*}$, in Step 4b. This differs from how δ^* was incorporated into OCBA by Chen and Kelton (2005), but $\mathcal{OCBA}_{\delta^*}$ was found to be more efficient in Branke et al. (2005).

All OCBA variations above were implemented as fully sequential procedures here ($q = 1$ and $\tau = 1$).

The OCBA heuristics in the literature to date assume normally distributed samples. Nonetheless, only the asymptotic normality of the posterior distribution of the mean is needed to justify the OCBA heuristic. In fact, posterior distributions converge to a normal distribution under relatively general sampling assumptions (Bernardo and Smith 1994, Theorem 5.14). That result can be used to asymptotically justify the OCBA heuristic under those relatively general sampling assumptions too, not just for normally distributed samples.

1.5 Summary of Tested Procedures

In summary, IZ procedures allocate samples in order to provide frequentist ($\mathrm{PCS}_{\mathrm{IZ}}$ or $\mathrm{EOC}_{\mathrm{IZ}}$) correct selection guarantees, but do not yet have provable properties regarding the EVI of those samples, nor do they quantify the posterior PCS given the output of a single application of a selection procedure. The VIP and OCBA procedures can quantify the posterior evidence for correct selection, and can stop when desired levels of evidence are achieved, but do not yet have provable frequentist correct selection guarantees. In addition, the VIP allocates samples in a way that provides provable statements about the EVI of those samples.

In addition to $\mathcal{KN}{+}{+}$, we tested six different allocation procedures, namely

- Equal, which allocates an equal number of samples to each alternative,
- two VIP procedures that allocate with a PCS (denoted 0-1) or EOC (denoted $\mathcal{LL}$) criterion,
- three OCBA procedures that allocate with a PCS (denoted $\mathcal{OCBA}$), PGS (denoted $\mathcal{OCBA}_{\delta^*}$), or EOC (denoted $\mathcal{OCBA}_{\mathrm{LL}}$) criterion.

Each allocation except for $\mathcal{KN}{+}{+}$ was used in combination with each of three stopping rules defined in §1.1 ($\mathcal{S}$, $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$, and $\mathrm{EOC}_{\mathrm{Bonf}}$). Overall, this resulted in 19 different procedures. So many variations were tested (a) to be inclusive and match all combinations in order to better understand the relative influence of each, (b) to unify separate streams of literature where small numbers of variants are compared at a time and numerical tests do not tend to be comparable, and (c) show the improvement in both VIP and OCBA procedures with stopping rules other than $\mathcal{S}$ (the default in all past VIP and OCBA work). Branke et al. (2005) reports preliminary results for other allocations and stopping rules that turned out to be less effective than those considered in this paper. We also tested the effect of including prior information about the means and variances in the VIP and OCBA configurations, as discussed in §3 below.

At each iteration, the time to compute an allocation is proportional to the square of the number of non-eliminated systems for $\mathcal{KN}{+}{+}$, to k^2 in the worst case for the VIP (to k if τ is large); and to k for the OCBA. For most practical applications, the time to compute the allocation is much shorter than the duration of a typical simulation. Each procedure can allocate multiple samples at a time if that is not the case.

2 Evaluation Criteria

There are several ways to evaluate selection procedures, including the theoretical, empirical, and practical perspectives. §1 indicates that the three approaches make different basic assumptions. Each uses different approximations or bounds. Theory that directly relates the three approaches is therefore difficult to develop, although theoretical derivations in Appendix B explain why $\mathcal{OCBA}_{\rm LL}$ and $\mathcal{LL}$ perform similarly.

We turn to the empirical and practical perspectives. The efficiency of a procedure is a frequentist measure of evidence for correct selection ($\rm PCS_{IZ}$, $\rm PGS_{IZ,\delta^*}$ or $\rm EOC_{IZ}$) as a function of the average number of replications, $\mathrm{E}[N]$. As a function of each problem instance and sampling allocation, the stopping rule parameters *implicitly* define *efficiency curves* in the $(\mathrm{E}[N], \rm PCS_{IZ})$ plane. For $\mathcal{KN}$++, or for the $\rm PGS_{Slep,\delta^*}$ stopping rule, for example, varying α^* generates an efficiency curve for any fixed δ^*. Varying the budget, B for the $\mathcal{S}$ stopping rule and varying β^* for the $\rm EOC_{Bonf}$ stopping rule also define efficiency curves. Efficiency curves for $\rm EOC_{IZ}$ and $\rm PGS_{IZ,\delta^*}$ are defined similarly. Appendix A.7 discusses the distribution of the number of samples that are required by each procedure, not just the mean.

Dai (1996) proved exponential convergence for ordinal comparisons in certain conditions, so efficiency curves might be anticipated to be roughly linear on a semi-log scale, $(\mathrm{E}[N], \log(1-\rm PCS_{IZ}))$. 'More efficient' procedures have lower efficiency curves.

Efficiency curves ignore the question of how to set a procedure's parameters to achieve a particular $\rm PCS_{IZ}$ or $\rm EOC_{IZ}$. As a practical matter, one expects some deviation between a stopping rule target, say $\mathrm{PCS} \geq 1-\alpha^* > 1/k$, and the actual $\rm PCS_{IZ}$ achieved. The deviation between the desired and realized performance is measured with *target curves* that plot $(\log\alpha^*, \log(1-\rm PCS_{IZ}))$ for PCS-based targets $1-\alpha^*$, and $(\log\beta^*, \log \rm EOC_{IZ})$ for opportunity cost targets β^*. Procedures whose target curves follow the diagonal $y = x$ over a range of problems are 'controllable' in that it is possible to set parameter values to obtain a desired level of correct selection. 'Conservative' procedures have target curves that tend to be below $y = x$, and are said to 'overdeliver' because the frequentist measure for correct selection exceeds the desired target. A procedure that is conservative may have a desirable frequentist guarantee for $\rm PCS_{IZ}$, but strong overdelivery results in excessive sampling. A controllable procedure may not have a $\rm PCS_{IZ}$ guarantee if the target curve varies slightly above and below the diagonal. We say that a procedure is *highly effective*, if it is both efficient and controllable.

Target curves can also assess whether Bayesian PCS goals map well to $\rm PCS_{IZ}$ or not (the VIP and OCBA do not yet claim $\rm PCS_{IZ}$ guarantees).

3 Test Bed Structure

A large number of problem instances assessed the strengths and weaknesses of each procedure. We literally explored many thousands of combinations of the number of systems, the first stage sampling size, specific configurations, allocations, stopping rules and their parameters, performance measures, etc. We tested random problem instances and the ability to use prior information about the unknown means. Design settings were chosen to explore first-order effects of the stopping rules and each parameter of the configurations and allocations, but not interactions. Appendix D gives further details about the structure of the experiments.

In a **slippage configuration (SC)**, the means of all systems except the best are tied for second best. A

SC is identified by the number of systems, the difference in means of the best and each other system, and the variances of each system. The parameters δ, ρ describe the configurations we tested.

$$\begin{aligned} X_{1j} &\overset{iid}{\sim} \texttt{Normal}\left(0, \sigma_1^2\right) \\ X_{ij} &\overset{iid}{\sim} \texttt{Normal}\left(-\delta, \sigma_1^2/\rho\right) \text{ for systems } i = 2, \ldots, k \end{aligned}$$

If $\rho = 1$, then all systems have the same variance, and $\rho < 1$ means that the best system has a smaller variance. We set $\sigma_1^2 = 2\rho/(1+\rho)$ so that $\mathrm{Var}[X_{1j} - X_{ij}]$ is constant for all $\rho > 0$.

In a **monotone decreasing means (MDM)** configuration, the means of all systems are equally spaced out so that some systems are quite a bit inferior to the best. The parameters δ and ρ describe the configurations that we tested. The outputs were jointly independent, and we set σ_1^2 like in SC.

$$X_{ij} \sim \texttt{Normal}\left(-(i-1)\delta, \sigma_1^2/\rho^{i-1}\right) \text{ for systems } i = 1, \ldots, k$$

Values of $\rho < 1$ mean that better systems have a smaller variance. For sufficiently small ρ, the probability that the worst system has the best observed mean is higher than the probability for the second best system.

For the SC and MDM configurations we tested hundreds, but not all, of the following combinations: $k \in \{2, 5, 10, 20, 50\}$, $\delta \in \{0.25, 0.354, 0.5, 0.707, 1\}$, $\rho \in \{0.125, 0.177, 0.25, \ldots, 2.828, 4\}$ (ratios of $\sqrt{2}$), and $n_0 \in \{4, 6, 10\}$. We then varied $\delta^* \in \{0, 0.05, 0.1, \ldots, 0.6\}$ and $\alpha^* \in [0.001, 0.5]$ for the $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ stopping rule and $\mathcal{KN}$++, and varied $\beta^* \in [0.001, 0.5]$ for the $\mathrm{EOC}_{\mathrm{Bonf}}$ stopping rule.

Assessments of selection procedures in the literature usually apply procedures to a specific set of structured problems, as above. Such structured problem instances are atypical in practice. A problem found in practice may be considered "random". **Random problem instance (RPI)** configurations sample the problem instance χ prior to applying a selection procedure. In each RPI configuration below, the output is jointly independent, $X_{ij} \overset{iid}{\sim} \texttt{Normal}\left(w_i, \sigma_i^2\right)$ for $i = 1, \ldots, k$, conditional on the problem instance. There is no objectively best distribution for sampling problem instances, so we make arbitrary choices and identify the biases of each.

The first RPI configuration (**RPI1**) samples χ from the normal-inverse gamma family. A random χ is generated by sampling the σ_i^2 independently, then sampling the W_i, given σ_i^2,

$$\begin{aligned} p(\sigma_i^2) &\sim \texttt{InvGamma}\left(b, c\right) \\ p(W_i \mid \sigma_i^2) &\sim \texttt{Normal}\left(\mu_0, \sigma_i^2/\eta\right). \end{aligned} \tag{10}$$

If $S \sim \texttt{InvGamma}\left(b, c\right)$, then $\mathrm{E}[S] = c/(b-1)$, $S^{-1} \sim \texttt{Gamma}\left(b, c\right)$, $\mathrm{E}[S^{-1}] = bc^{-1}$ and $\mathrm{Var}[S^{-1}] = bc^{-2}$. Increasing η makes the means more similar and therefore the problem harder. We set $c = b - 1 > 0$ to standardize the mean of the variance to be 1, and set $\mu_0 = 0$. Increasing b reduces the difference between the variances. We tested many combinations out of $k \in \{2, 5, 10, 20\}$, $\eta \in \{0.354, 0.5, \ldots, 4\}$ (ratios of $\sqrt{2}$), $b \in \{2.5, 100\}$, and $n_0 \in \{4, 6, 10\}$. The derivations of the VIP and OCBA procedures assume $\eta \to 0$.

The RPI1 configuration permits a test of whether the VIP and OCBA procedures can benefit from using the sampling distribution of χ in (10) to describe prior judgement about the means and variances of each system. §1 does not allow for this directly, but the mathematical development to do so was provided elsewhere for the VIP (Chick and Inoue 1998, 2001a). In summary, the posterior distribution of W_i, given the prior distribution in (10) and data $\mathcal{E}_i = (x_{i1}, \ldots, x_{in_i})$, is

$$\begin{aligned} p(\sigma_i^2 \,|\, \mathcal{E}_i) &\sim \texttt{InvGamma}\left(b', c'\right) \\ p(W_i \,|\, \sigma_i^2, \mathcal{E}_i) &\sim \texttt{Normal}\left(\mu_0', \sigma_i^2/\eta'\right), \end{aligned}$$

where $b' = b + n_i/2$, $c' = c + (\frac{\eta n_i}{\eta+n_i}(\mu_0 - \bar{x}_i)^2 + \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2)/2$, $\mu_0' = \frac{\eta\mu_0+n_i\bar{x}_i}{\eta+n_i}$, and $\eta' = \eta + n_i$. To apply that result to all VIP procedures in §1.3, substitute each $\bar{x}_i$ with μ_0'; replace each $\hat{\sigma}_i^2$ with c'/b'; and replace each n_i with η', except in the degrees of freedom, where $n_i - 1$ should be replaced with $2b'$. To date, the $\mathcal{OCBA}$ has always assumed a noninformative prior distribution. Nonetheless, analogous substitutions allow $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\mathrm{LL}}$ to use other prior distributions for the unknown means and variances.

A second RPI configuration (**RPI2**) samples problem instances from a distribution other than normal-inverted gamma to reduce any potential advantage for the VIP and OCBA approaches. We chose RPI2 to independently sample from:

$$\begin{aligned} p(\sigma_i^2) &\sim \texttt{InvGamma}\left(b, c\right) \\ p(W_i \,|\, \sigma_i^2) &\sim (-1)^a\texttt{Exponential}\left((\eta/\sigma_i^2)^{1/2}\right), \end{aligned}$$

where the mean of an $\texttt{Exponential}\,(\lambda)$ distribution is $1/\lambda$. There are typically several competitors for the best if $a = 1$ and few competitors for the best if $a = 0$. A larger η makes for harder problems with closer means. Heterogeneity in the variances is controlled with b and c. We tested values of k, η, b, n_0 as in RPI1.

The SC favors IZ procedures in that IZ procedures provide a minimal target performance with respect to a least favorable configuration (LFC). For many IZ procedures, the SC with $\delta = \delta^*$ is a LFC. Although the LFC of the $\mathcal{KN}$ family has not been proven, empirical studies of $\mathcal{KN}$-type procedures often assess that configuration. The RPI1 (η near 0) favors the VIP and OCBA, as the derivation of those procedures assumes prior probability models that are similar to the sampling distribution of the problem instances. The MDM, RPI1 (larger η) and RPI2 configurations do not appear to favor any procedure in this paper.

4 Empirical Results

It is impossible to present all of the numerical results in one short article. This section summarizes the main qualitative observations from the analysis. Appendix A provides more numerical results. Appendix C describes the implementation.

In the following analysis, each point that defines each efficiency and target curve (one for each combination of problem instance, procedure, and choice of parameters) was estimated with 10^5 samples (applications of a procedure). To sharpen the contrasts between different procedures, common random numbers (CRN) were used to generate common RPI configurations, and to synchronize the samples observed across procedures. Samples were independent within each procedure. The notation $\mathcal{KN}{+}{+}_{\delta^*}$ specifies the choice of δ^* for $\mathcal{KN}{+}{+}$. By default, $n_0 = 6$ unless specified otherwise.

Two Systems, SC/MDM. When there are $k = 2$ systems, the SC and MDM configurations are equivalent, and $\mathrm{PICS}_{\mathrm{IZ}}$ is proportional to $\mathrm{EOC}_{\mathrm{IZ}}$. When the variances are known and equal, it is optimal to sample

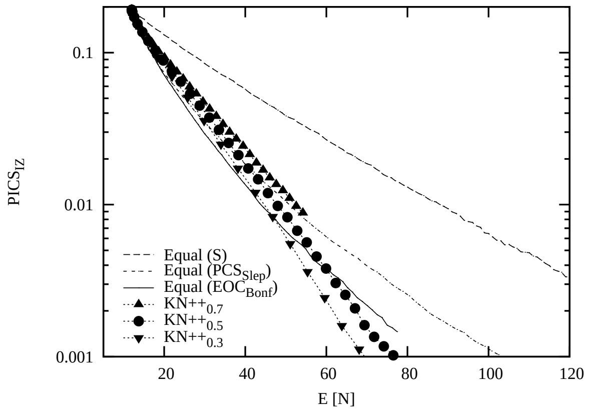


Figure 1: *Efficiency* of different stopping rules (for Equal) and $\mathcal{KN}$++ (SC, $k = 2$, $\delta = 0.5$, $\rho = 1$).

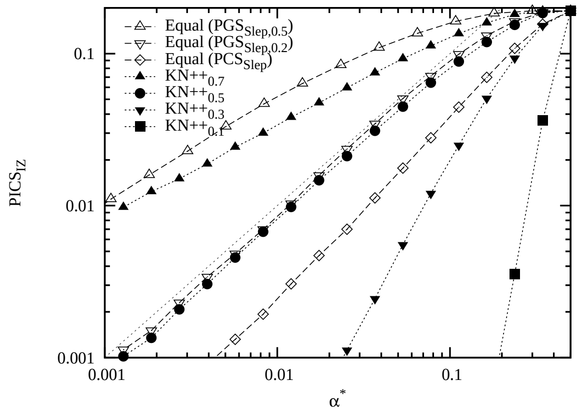


Figure 2: *Target* curves of Equal($\text{PGS}_{\text{Slep},\delta^*}$) and $\mathcal{KN}$++ depend on δ^*. (SC, $k = 2$, $\delta = 0.5$, $\rho = 1$).

equally often from both systems from both frequentist and Bayesian EVI perspectives (for both the $0 - 1$ and EOC loss functions, e.g. Gupta and Miescke 1994), so Procedure Equal samples optimally for such configurations.

Figure 1 demonstrates the effect of different stopping rules on efficiency for a given sampling allocation (here, Equal). For $k = 2$ in particular, $\mathcal{KN}$++ samples each system equally often until the stopping criterion is met, so it is equivalent to the Equal allocation with a special stopping rule. The EOC-based stopping rule is more efficient than the PCS-based stopping rule. Both are much more efficient than stopping after a fixed budget ($\mathcal{S}$) because any additional sampling is adapted to the level of evidence observed so far. That relative ordering of the stopping rules (EOC_{Bonf} beats PCS_{Slep}, which beats $\mathcal{S}$) was observed for *all* VIP and OCBA allocations with a similar order of magnitude difference. Similar effects were seen for different δ.

Figure 1 also illustrates that the efficiency of $\mathcal{KN}$++ depends on the setting of δ^*, an observation that holds for *all* configurations that we checked for *all* procedures that use δ^* (see also Figures 20 and 30 in the Online Companion). In addition, we note that the efficiency curves for $\mathcal{KN}$++ and Equal($\mathcal{S}$) are straighter than for the PCS or EOC stopping rules. Those Bayesian stopping rules cause a slight curvature that is linked to the choice of n_0 (a smaller n_0 gives more curvature, more on n_0 below). For a higher PICS, Equal(EOC_{Bonf}) is more efficient than $\mathcal{KN}$++, while for a very low PICS, $\mathcal{KN}$++ beats Equal(EOC_{Bonf}).

A target plot can show by how much a procedure deviates from the desired goal (overdelivery and underdelivery). The SC configuration for the target plot in Figure 2 is the same configuration that was used for the efficiency plot in Figure 1. The target line for the PCS-based stopping rule is below the diagonal, which means that the obtained PICS_{IZ} is smaller than the desired goal α^*. For example, for a desired $\alpha^* = 0.02$, a $\text{PICS}_{\text{IZ}} = 0.005$ is obtained with Equal(PICS_{IZ}). As δ^* increases for the $\text{PGS}_{\text{IZ},\delta^*}$ stopping rule, the target curve shifts upward (the target is not obtained for $\delta^* > 0.5$). The target plot for $\mathcal{KN}$++$_{0.5}$ follows the diagonal well, meaning that the obtained PICS_{IZ} matches the desired goal α^* well for this configuration (this was observed for $\mathcal{KN}$++ with all SC configurations that had $k \geq 2$ and $\delta^* = \delta$). As δ^* is reduced, the line tilts downward, meaning that $\mathcal{KN}$++ becomes extremely conservative.

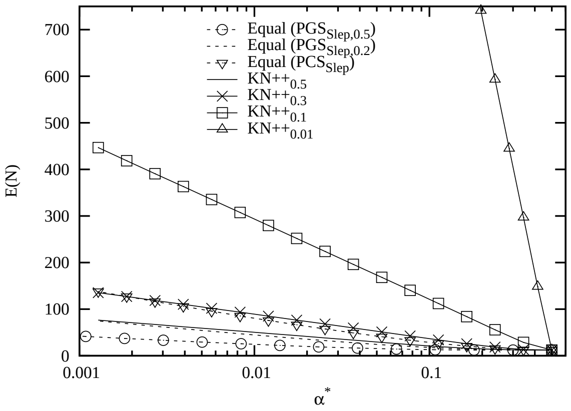


Figure 3: *Sensitivity* of $\mathrm{E}[N]$ to choice of δ^*. (SC, $k = 2$, $\delta = 0.5$, $\rho = 1$).

The consequences of conservativeness can be derived by looking at the efficiency and target plots together. For example, the x-axis intercept for the target plot of $\mathcal{KN}{+}{+}_{0.3}$ is 0.025, meaning that a goal of $\alpha^* = 0.025$ resulted in an actual $\mathrm{PICS}_{\mathrm{IZ}} = 0.001$ (overdelivers PCS). In Figure 1, the mean number of replications for $\mathcal{KN}{+}{+}_{0.3}$ that corresponds to $\mathrm{PICS}_{\mathrm{IZ}} = 0.001$ is $\mathrm{E}[N] = 68$ replications. Figure 1 also indicates that the mean number of replications that are required to achieve a desired $\mathrm{PICS}_{\mathrm{IZ}}$ of $\alpha^* = 0.025$ with $\mathcal{KN}{+}{+}_{0.3}$ is $\mathrm{E}[N] = 34$, i.e. the procedure runs twice as long as necessary.

Figure 3 illustrates this mapping from α^* to $\mathrm{E}[N]$ more fully. $\mathcal{KN}{+}{+}$ is extremely conservative when $\delta^* \ll \delta$, but the $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ stopping rule is not (here, with Equal). This shows that a procedure with good efficiency curves can require *much* more sampling than is required to achieve a given level of evidence for correct selection, if that procedure is conservative. On the other hand, the PGS stopping rule has the risk of delivering a $\mathrm{PICS}_{\mathrm{IZ}}$ higher than desired (for Equal($\mathrm{PGS}_{\mathrm{Slep},0.5}$), the PCS target is not met).

Another aid to interpreting efficiency and target curves comes from noting that the efficiency curve for $\mathcal{KN}{+}{+}_{0.7}$ in Figure 1 terminates near the level of $\mathrm{PICS}_{\mathrm{IZ}} = 0.009$. That is symptomatic of an underdelivery of $\mathrm{PICS}_{\mathrm{IZ}}$, as we tested values of α^* down to 0.001. Figure 2 corroborates this underdelivery, as the target plot for $\mathcal{KN}{+}{+}_{0.7}$ in Figure 2 is above the diagonal. This particular case of underdelivery is fully consistent with the IZ approach, since the indifference zone parameter exceeds the difference in means ($\delta^* = 0.7 > \delta = 0.5$).

Figure 4 shows how different sampling allocations compare for a given stopping rule (here, for $\mathrm{EOC}_{\mathrm{Bonf}}$). Equal performs most efficiently (it is optimal for this particular setting), with 0-1 and $\mathcal{OCBA}$ following. $\mathcal{LL}$ performs identically to 0-1 for this problem (not shown, to avoid cluttering the figure). A similar precedence is observed for the $\mathrm{PCS}_{\mathrm{Slep}}$ and $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ stopping rules. For the $\mathcal{S}$ stopping rule, all VIP and OCBA allocations perform about the same as Equal (not shown, to avoid clutter).

With adaptive stopping rules ($\mathrm{EOC}_{\mathrm{Bonf}}$, $\mathrm{PCS}_{\mathrm{Slep}}$, $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$), a large number of initial samples per system, n_0, limits the opportunity to make an early selection, but a small n_0 increases the probability of poor estimates of the output mean and variance. For the posterior marginal distributions of the unknown means to have a finite variance, we require $n_0 \geq 4$. Figure 5 shows that increasing n_0 in Procedure Equal($\mathrm{EOC}_{\mathrm{Bonf}}$) increases the number of samples required to reach relatively low levels of evidence for correct selection, but increases

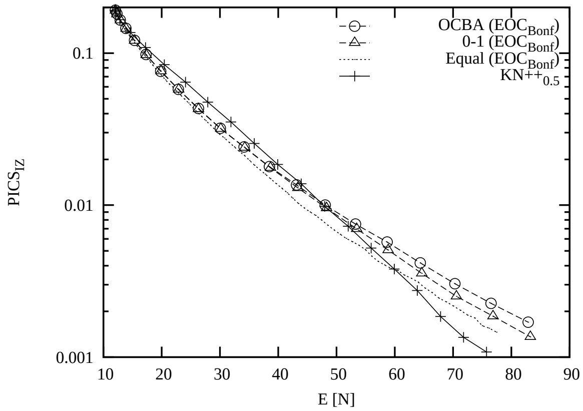


Figure 4: Efficiency of different allocation procedures (SC, $k = 2$, $\delta = 0.5$, $\rho = 1$).

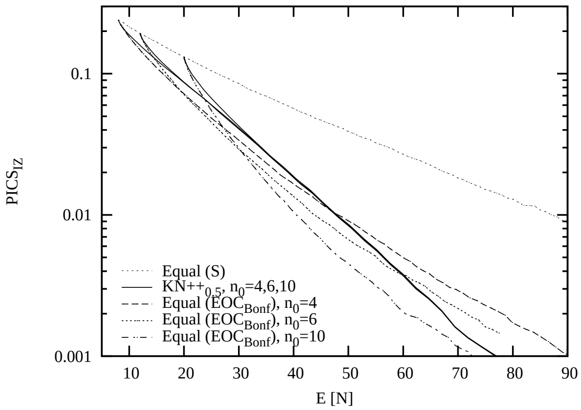


Figure 5: Influence of n_0 (SC, $k = 2$, $\delta = 0.5$, $\rho = 1$).

the efficiency of the procedure to reach high levels of evidence for correct selection. The differences in the curves are predominantly attributed to output that causes sampling to stop after very few samples due to misleadingly low variance and PICS estimates. The OCBA and VIP procedures behave similar to Equal in this respect for each stopping rule. With the nonadaptive stopping rule ($\mathcal{S}$), they seem insensitive to n_0. Figure 5 also shows that $\mathcal{KN}{+}{+}$ is insensitive to n_0, an observation that held in general.

The EOC$_{\rm Bonf}$ stopping rule is sensitive to the difference between the two systems, δ (see Figure 6). It slightly underdelivers EOC$_{\rm IZ}$ for small δ, and significantly overdelivers for large δ.

Overall, for SC with $k = 2$ and common variances ($\rho = 1$), Equal(EOC$_{\rm Bonf}$) and $\mathcal{KN}{+}{+}$ are the most efficient. No procedure is fully controllable for SC, $k = 2$. The Online Companion also argues that we could not find a general way to "trick" the procedure by setting δ^* and α^* to nontraditional values in order to achieve some actually desired level of PCS$_{\rm IZ}$.

The remarks so far presume a common variance ($\rho = 1$). When $\rho \neq 1$, the equal allocation is not optimal, and more samples should be distributed to the system with a higher variance. We observed that $\mathcal{KN}{+}{+}$ and Equal become less efficient than the Bayesian allocations as ρ is changed away from 1 (Online Appendix).

SC with $k > 2$ systems. It is not optimal to sample each system equally often if $k > 2$, so $\mathcal{KN}{+}{+}$ and Equal are no longer optimal in this setting, even if the variances are equal.

A comparison of the efficiency of the different allocation rules, for the EOC$_{\rm Bonf}$ stopping rule and $k = 10$ systems, is illustrated in Figure 7. The settings are comparable to Figure 4, which had $k = 2$ systems. While Equal is optimal for $k = 2$, it performs worst for $k = 10$. The most efficient allocations are $\mathcal{OCBA}_{\rm LL}$ and $\mathcal{OCBA}$, then $\mathcal{LL}$ closely behind. $\mathcal{KN}{+}{+}$ is much less efficient than $\mathcal{OCBA}_{\rm LL}$, $\mathcal{OCBA}$ and $\mathcal{LL}$ (each with EOC$_{\rm Bonf}$ as a stopping rule). The difference between the efficiency of the Bayesian procedures relative to the Equal and $\mathcal{KN}{+}{+}$ procedures increases with k (tested $k = 2, 5, 10, 20$), presumably because they can allocate more samples to the most relevant (best) system from the beginning. The qualitative nature of this claim does not change as δ and ρ are individually varied from the values used for the plot.

Other observations for $k = 2$ that also hold for $k = 5, 10$ and 20 include: the precedence of the effectiveness of stopping rules (EOC$_{\rm Bonf}$ beats PCS$_{\rm Slep}$ which beats $\mathcal{S}$); the importance of a sufficiently large

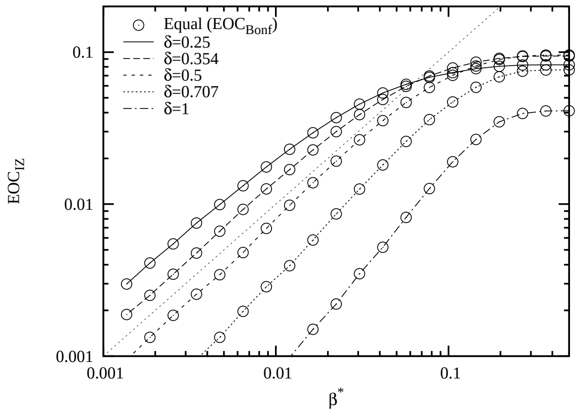


Figure 6: Influence of δ on target performance (SC, $k = 2$, $\rho = 1$).

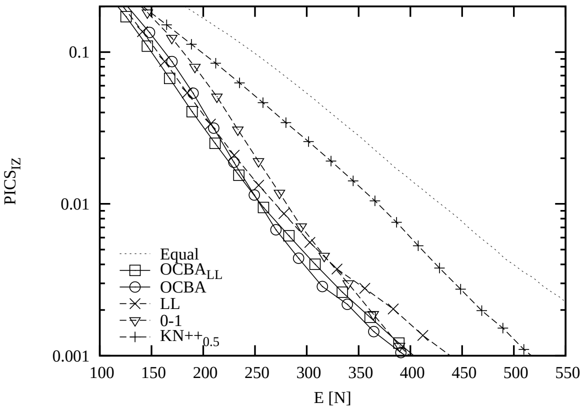


Figure 7: Efficiency with EOC$_{\text{Bonf}}$ stopping rule (SC, $k = 10$, $\delta = 0.5$, $\rho = 1$).

n_0 for the VIP and OCBA procedures; the sensitivity of $\mathcal{KN}$++ to δ^* but not n_0.

Monotone Decreasing Means add the complication that EOC$_{\text{IZ}}$ is not proportional to PCS$_{\text{IZ}}$ when $k > 2$. Figure 8 (which has $k = 100$), and Figure 26 of the Online Companion, illustrate that the EOC$_{\text{Bonf}}$ stopping rule outperforms PCS$_{\text{Slep}}$, which beats $\mathcal{S}$, an order observed for the Equal and *all* VIP and OCBA allocations, and *all* MDM configurations tested. The EOC$_{\text{Bonf}}$ stopping rule outperforms PCS$_{\text{Slep}}$ not only for EOC$_{\text{IZ}}$ efficiency, but also for PCS$_{\text{IZ}}$ efficiency.

Figure 8 illustrates another key observation for the MDM configurations. Procedure $\mathcal{KN}$++ with $\delta^* = \delta$ is typically more efficient than the original VIP and OCBA procedures, which use the $\mathcal{S}$ stopping rule. However, the VIP and OCBA allocations with the new EOC$_{\text{Bonf}}$ stopping rule are more efficient than $\mathcal{KN}$++, due to the flexibility of EOC$_{\text{Bonf}}$ to stop when, and only when, sufficient evidence for correct selection is obtained. For each MDM configuration tested, $\mathcal{LL}$(EOC$_{\text{Bonf}}$) and $\mathcal{OCBA}_{\text{LL}}$(EOC$_{\text{Bonf}}$) are not statistically different for efficiency, and they were the most efficient allocations, as in Figure 8. Those two procedures also perform roughly similar in target plots for most configurations. $\mathcal{OCBA}_{\delta^*}$ and 0-1 performed similar to those procedures when EOC$_{\text{Bonf}}$ was used. With the $\mathcal{S}$ stopping rule, however, 0-1 performed poorly.

Figure 9 illustrates that the target performance of $\mathcal{KN}$++ is again very sensitive to the parameter δ^*. While $\mathcal{KN}$++ adheres to the target quite well for SC when $\delta^* = \delta$, it significantly overdelivers even for this setting for the MDM configuration. Procedure $\mathcal{OCBA}_{\text{LL}}$(PGS$_{\text{Slep},\delta^*}$) is also sensitive to the parameter δ^*, and fails to obtain the desired PICS$_{\text{IZ}}$ for $\delta^* > 0.2$ even though $\delta = 0.5$. While the target curves for $\mathcal{OCBA}_{\text{LL}}$(PGS$_{\text{Slep},\delta^*}$) shift roughly parallel to the diagonal, the target curves of $\mathcal{KN}$++ change in slope. This means that $\mathcal{KN}$++ becomes more conservative as more extreme levels of evidence (α^*) are sought, if $\delta^* < \delta$, but $\mathcal{OCBA}_{\text{LL}}$(PGS$_{\text{Slep},\delta^*}$) tends to be conservative by the same amount, on a log scale. Overall, as for SC, $\mathcal{KN}$++ and PGS$_{\text{Slep},\delta^*}$ are very sensitive to δ^* and thus not controllable in the sense defined above.

Figure 10 illustrates the effect of the output variance ratio ρ on different procedures for MDM with $k = 10$. With $\rho = 1$ (equal variance), the best PCS$_{\text{Slep}}$ procedures perform somewhat more efficiently than $\mathcal{KN}$++. Increasing ρ (best systems have larger variance) has little effect on the relative performance of the procedures. Decreasing ρ to 0.5 (very large variance for the worst systems) increases the total number of

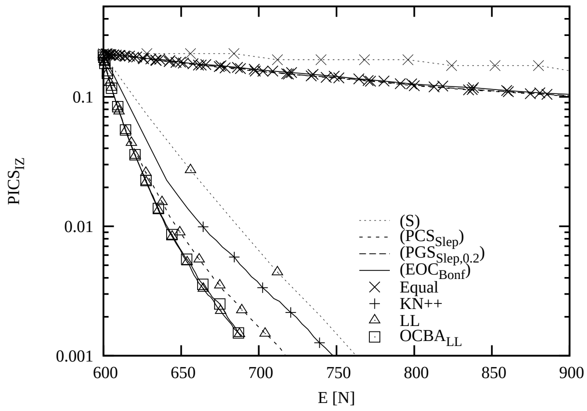


Figure 8: Different stopping rules (line types) and allocation procedures (symbols) (MDM, $k = 100$, $\delta = 0.5$, $\rho = 1$).

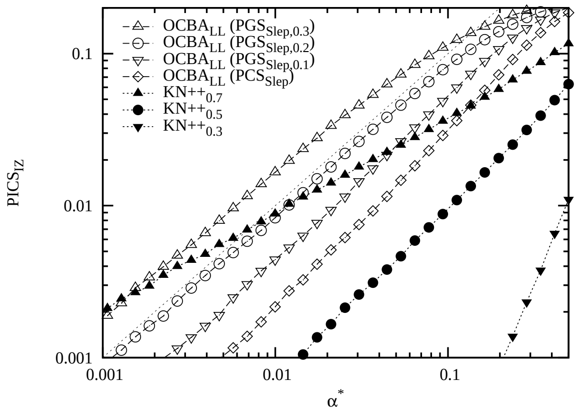


Figure 9: Target graphs for $\mathcal{KN}{++}$ and 0-1 (MDM, $k = 10$, $\delta = 0.5$, $\rho = 1$).

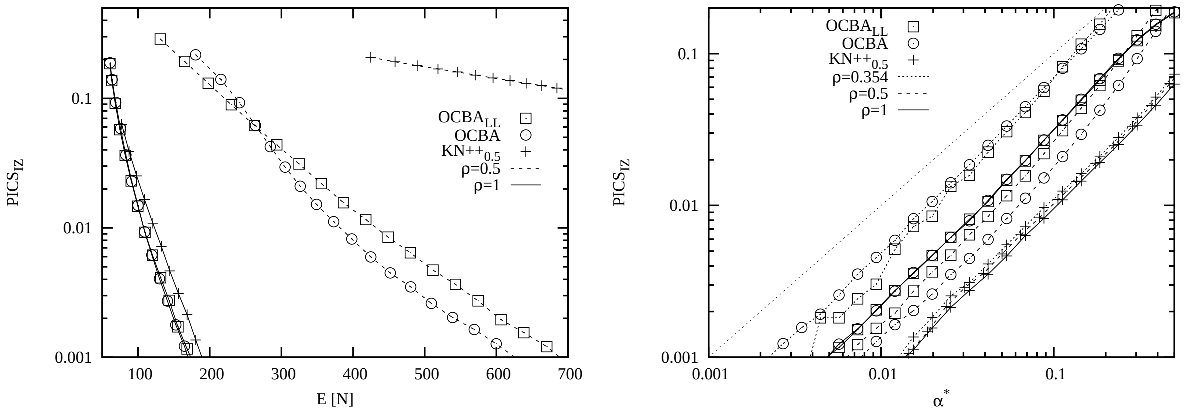


Figure 10: Effect of variance ratio ρ on efficiency (left panel) and target performance (right panel). (MDM, $k = 10$, $\delta = 0.5$). Procedures $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\text{LL}}$ use PCS$_{\text{Slep}}$ stopping rule. Procedures $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ perform the same (not shown to avoid clutter).

samples for all procedures, but particularly deteriorates the efficiency of $\mathcal{KN}{++}$ (in far right of efficiency plot) relative to the PCS$_{\text{Slep}}$ procedures. $\mathcal{KN}{++}$ also overdelivers more than some other procedures (right panel). The reason is that $\mathcal{KN}{++}$ samples all noneliminated systems, even those whose means have been estimated with high precision, until a system is selected as best. The target curves for all procedures are relatively insensitive to ρ (the target curve depends primarily on the difference between the two systems competing most for best, so efficiency is affected more than the target curve).

For all SC and MDM configurations and almost all sampling procedures, the efficiency curves exhibit some curvature. We found several explanations for curved efficiency lines for the OCBA and VIP procedures. One, a small n_0 leads to poor initial estimates of the variance, with a potential for either (a) early stopping if PICS is strongly underestimated, or (b) a massive number of samples being required if an extremely low PICS or EOC is desired, initial estimates suggest that the best system is worst, and the procedure then tries to distinguish between the equal systems in the SC. Both cases are alleviated by increasing n_0. Two, the test bed

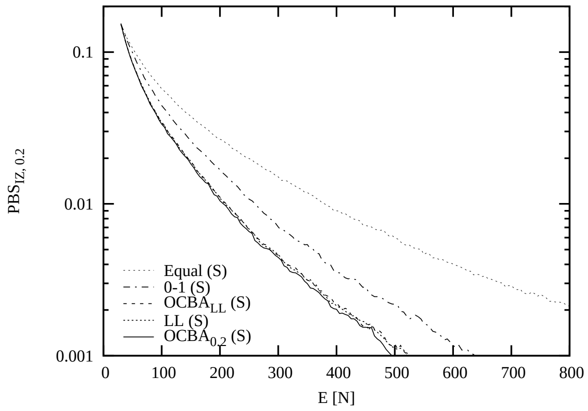


Figure 11: PBS$_{\delta^*}$ efficiency of sampling allocations with $\mathcal{S}$ stopping rule (RPI1, $k = 5$, $\eta = 1$, $b = 100$).

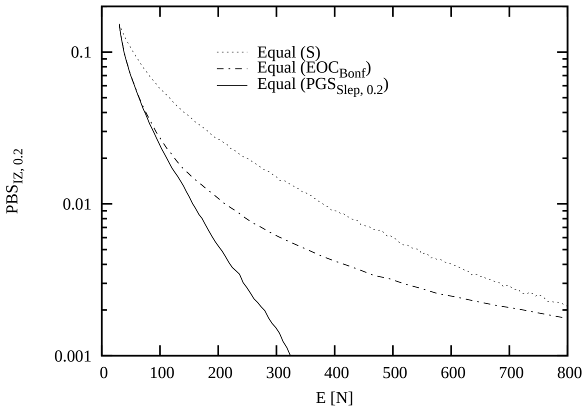


Figure 12: PBS$_{\delta^*}$ efficiency of stopping rules with Equal allocation, (RPI1, $k = 5$, $\eta = 1$, $b = 100$).

pushed the procedures to new limits for numerical stability. Preliminary efficiency plots for some procedures were somewhat more curved than those presented here. Appendix C describes computational techniques that we used to reduce that curvature. We believe that this cause was eliminated. Three, exponential convergence results for ordinal comparisons are asymptotic and available for only some procedures, so straight lines should not be expected at all levels of PICS$_{\mathrm{IZ}}$ and EOC$_{\mathrm{IZ}}$ for all procedures.

Random Problem Instances 1. For all RPI configurations, the metrics for the evidence for correct selection in the efficiency and target plots are generalized to be expectations over the sampling distribution of the problem instances, e.g. PCS$_{\mathrm{IZ}} = \mathrm{E}_{\boldsymbol{\chi}}[\mathrm{PCS}_{\mathrm{IZ}}(\boldsymbol{\chi})]$. One must choose $\delta^* > 0$ for the PGS$_{\mathrm{Slep},\delta^*}$ stopping rule because there is a reasonable probability that the two best systems have very similar means, in which case $\delta^* = 0$ results in excessive sampling (so $\delta^* = 0$ is to be avoided in practice). For the same reason, we measure efficiency by the probability of a bad selection, PBS$_{\mathrm{IZ},\delta^*}$, instead of PICS$_{\mathrm{IZ}}$, in this section. A bad selection has a mean that is at least δ^* worse than the mean of the best system.

Procedures that use an indifference zone parameter δ^* can conceivably have their efficiency measured with a PGS$_{\delta^{**}}$ that use an indifference zone value $\delta^{**} \neq \delta^*$, but philosophically this seems inconsistent. We therefore use the same "matching" indifference zone parameter δ^* for both allocations and empirical measurements, unless otherwise specified (see also Figure 30 in the Online Companion).

For basically *all* RPI settings, the $\mathcal{LL}$, $\mathcal{OCBA}_{\mathrm{LL}}$ and $\mathcal{OCBA}_{\delta^*}$ allocation rules are more or less equally efficient. The 0-1 allocation is generally less efficient (it is derived with more approximations, and wastes samples trying to distinguish between two very close competitors in the RPI) and Equal is worst. Figure 11 makes this point for the $\mathcal{S}$ stopping rule.

While the difference in efficiency among the allocation rules is rather small for RPI, the performance of the stopping rules varies widely. Figure 12 compares different stopping rules in combination with Equal allocation based on PGS$_{\mathrm{IZ},\delta^*}$ efficiency. Clearly, the PGS$_{\mathrm{Slep},\delta^*}$ stopping rule with a matching δ^* is the most efficient (and the target curve is also good in this setting).

For EOC$_{\mathrm{IZ}}$ efficiency, settings for δ^* exist so that PGS$_{\mathrm{Slep},\delta^*}$ is more efficient than the EOC$_{\mathrm{Bonf}}$ stopping rule, see Figure 13 (left panel). But we were not able to find a way to pick δ^* in PGS$_{\mathrm{Slep},\delta^*}$ to control EOC$_{\mathrm{IZ}}$

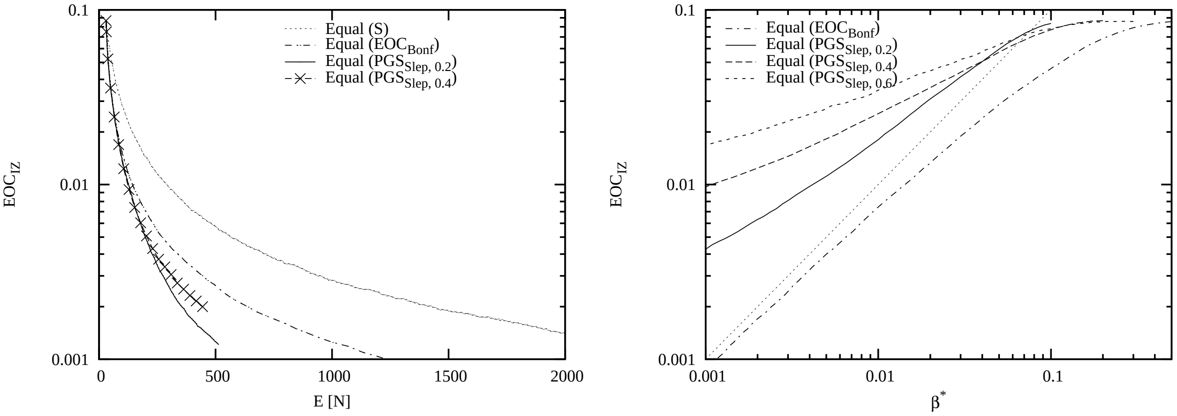


Figure 13: Efficiency (left) and target (right) for Equal allocation and different stopping rules w.r.t. EOC$_{\rm IZ}$ (RPI1, $k = 5$, $\eta = 1$, $b = 100$). For PGS$_{\text{Slep},\delta^*}$ stopping, β^* is approximated by $\alpha^*\delta^*$

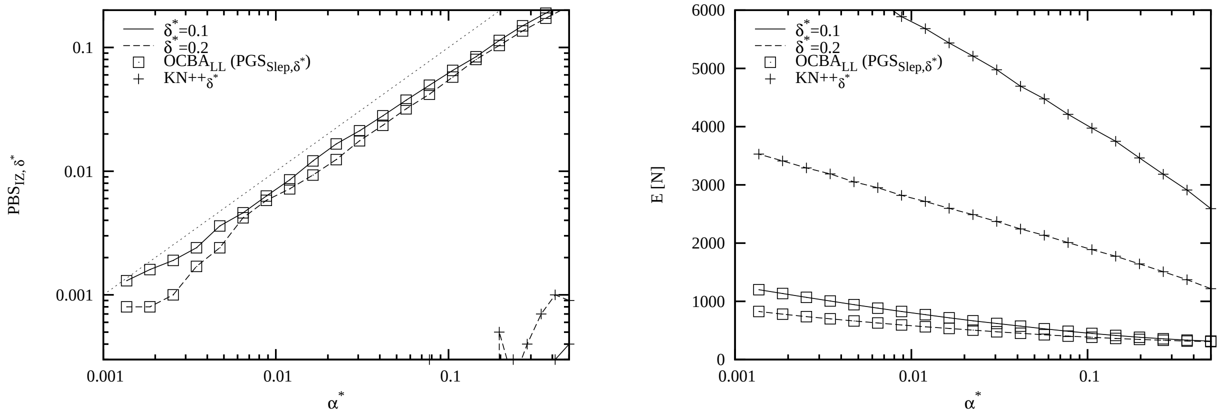


Figure 14: Comparison of $\mathcal{OCBA}_{\rm LL}$(PGS$_{\text{Slep},\delta^*}$) and $\mathcal{KN}$++ for PBS$_{\text{IZ},\delta^*}$ target (left panel) and number of samples (right panel) for different δ^* and α^* (RPI1, $k = 50$, $\eta = 1$, $b = 100$).

(the right panel shows a lack of controllability if one were to try to control EOC$_{\rm IZ}$ by setting $\beta^* = \delta^*\alpha^*$, which is the expected opportunity cost for the LFC with respect to PCS for a number of IZ procedures).

Several efficiency curves, in particular Equal(S) in Figure 12 and Figure 13, are more curved for the RPI1 results than for the SC and MDM results. That curvature is largely due to a very large number of samples for a few very "hard" configurations (the best two systems have very close means and large variances).

For the RPI1 configurations, $\mathcal{LL}$, $\mathcal{OCBA}_{\rm LL}$ and $\mathcal{OCBA}_{\delta^*}$, each with the PGS$_{\text{Slep},\delta^*}$ stopping rule, typically outperform $\mathcal{KN}$++ for efficiency and controllability. Figure 14 illustrates this observation by depicting the effect of different δ^* and α^* on the number of samples required by $\mathcal{KN}$++ and $\mathcal{OCBA}_{\rm LL}$(PGS$_{\text{Slep},\delta^*}$), and the ability to deliver the desired PGS (the δ^* parameter of the procedure is also used to measure PGS$_{\text{IZ},\delta^*}$). For this figure, with $k = 50$, the number of macroreplications was reduced to 10^4 to keep simulation time reasonable. The left panel shows that $\mathcal{OCBA}_{\rm LL}$(PGS$_{\text{Slep},\delta^*}$) follows the target much better than $\mathcal{KN}$++ (which is barely visible at the lower right of the figure). The right panel shows that the penalty for conservativeness is a significantly higher sampling effort for a given desired α^*. For example, for $\alpha^* = 0.01$ and $\delta^* = 0.1$, $\mathcal{OCBA}_{\rm LL}$(PGS$_{\text{Slep},\delta^*}$) requires 803 samples on average, while $\mathcal{KN}$++ requires 5845 samples.

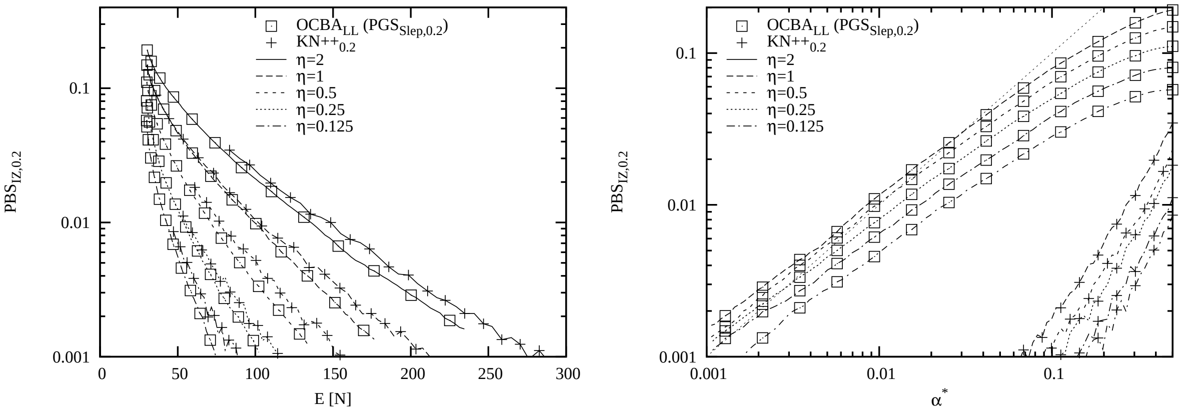


Figure 15: Influence of η on efficiency and target for PCS-based procedures (RPI1, $k = 5$, $b = 100$).

As can be seen in Figure 15, the sensitivity with respect to η in the RPI configurations is much smaller than the sensitivity with respect to δ observed for the SC and MDM configurations in Figure 6. Note that the difference between the best and second best system is proportional to η^{-2}. For efficiency (left panel), $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ slightly outperforms $\mathcal{KN}{+}{+}$. For the target plot, $\mathcal{KN}{+}{+}$ consistently and strongly overdelivers, while $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ meets the target rather well over all η tested. Also, procedures with the EOC$_{\text{Bonf}}$ stopping rule follow an EOC$_{\text{IZ}}$ target well (not shown for this setting, but the pattern is like that of Figure 15 for PBS$_{\text{IZ},\delta^*}$, and for EOC$_{\text{IZ}}$ in Figure 16 below).

Observations from the SC and MDM configurations that are also valid for RPI1 include: The Bayesian procedures are sensitive to n_0 (pick $n_0 \geq 6$, or even 10 if practical), while $\mathcal{KN}{+}{+}$ is not; $\mathcal{KN}{+}{+}$ becomes less efficient relative to the Bayesian procedures as k increases (Online Companion, Figures 31 and 32).

If prior knowledge on the distribution of means and variances is available, this can be integrated into the Bayesian procedures, as described in §3. The benefit of doing so, when possible, is apparent in Figure 16 (left panel). The top line shows the efficiency of the standard Equal allocation with the budget ($\mathcal{S}$) stopping rule. This can be improved stepwise by switching to a flexible allocation ($\mathcal{OCBA}_{\text{LL}}(\mathcal{S})$), then using an adaptive stopping rule ($\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$), and finally using prior information ($\mathcal{OCBA}^{\text{prior}}_{\text{LL}}(\text{EOC}^{\text{prior}}_{\text{Bonf}})$). These changes reduce the mean number of samples required to achieve a loss of 0.01 from 291 for Equal($\mathcal{S}$) to 164 for $\mathcal{OCBA}_{\text{LL}}(\mathcal{S})$, then to 94 for $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$, and finally to 79 for $\mathcal{OCBA}^{\text{prior}}_{\text{LL}}(\text{EOC}^{\text{prior}}_{\text{Bonf}})$. Controllability is only slightly affected by using prior information in this test (right panel).

Random Problem Instances 2. The RPI2 configuration samples random problem instances with a distribution that does not match the underlying assumptions of the derivations of the VIP and OCBA procedures. Still, Procedure $\mathcal{LL}$ and $\mathcal{OCBA}_{\text{LL}}$ again perform almost identically for efficiency and controllability, so only the latter is shown in plots.

Figure 17 compares the efficiency and target curves for $\mathcal{KN}{+}{+}$ and $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$. Procedures $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ are somewhat less efficient than $\mathcal{KN}{+}{+}$ if there are several good systems ($a = 1$, left panel). The difference is smaller for larger δ^*. Procedures $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ are more controllable than $\mathcal{KN}{+}{+}$, which is conservative and significantly overdelivers PBS$_{\text{IZ},0.2}$ (right panel). The RPI2 configuration with few good systems ($a = 0$) is similar to RPI1 with respect to the long tail distribution of the good systems,

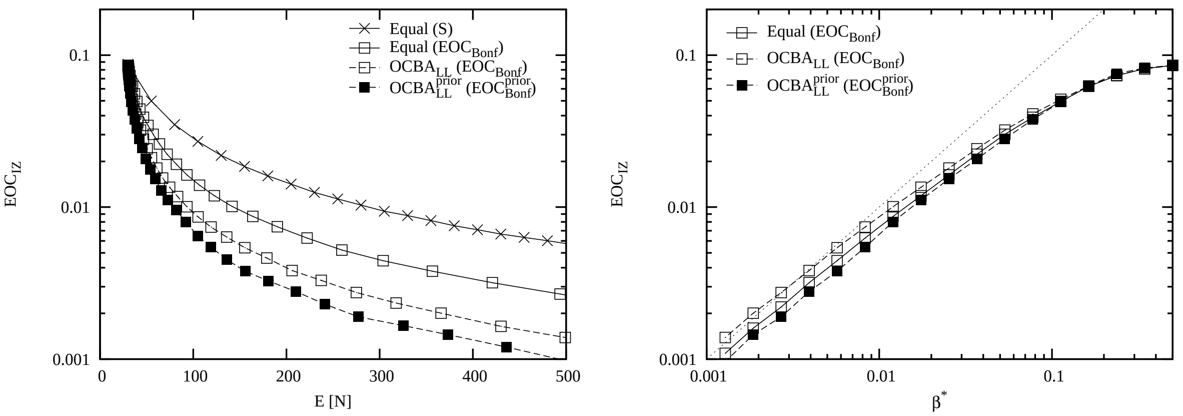


Figure 16: Effect of allocation, stopping rule and prior information (RPI1, $k = 5$, $\eta = 1$, $b = 100$).

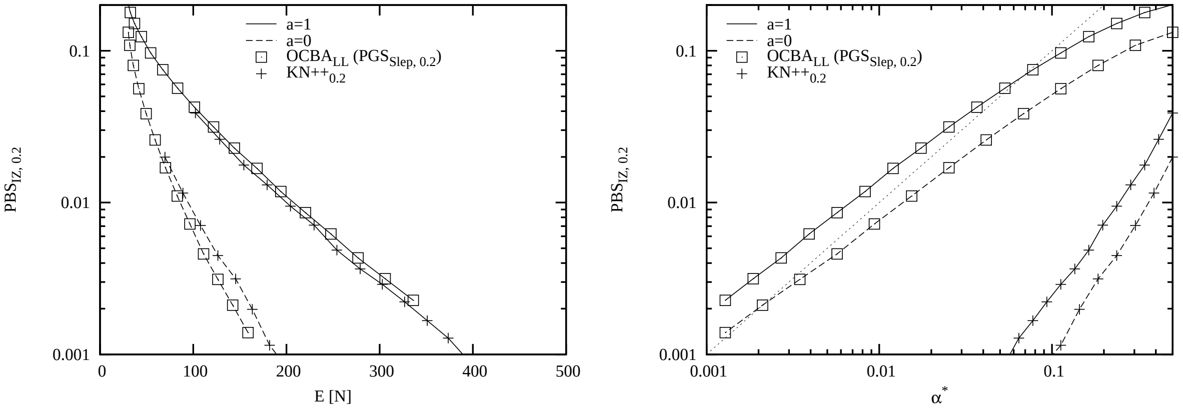


Figure 17: Efficiency and target for RPI2 ($k = 5$, $\eta = 1$, $b = 100$). $\mathcal{LL}$ and $\mathcal{OCBA}_{\mathrm{LL}}$ perform the same.

so it is not surprising that results are very similar. Appendix A.6 has more graphs.

On the whole, $\mathcal{OCBA}_{\mathrm{LL}}$ and $\mathcal{LL}$ perform very well for RPI2 even though the problem instances do not have the normal-inverted gamma distribution that is implicit in their derivation. A small degradation in efficiency relative to $\mathcal{KN}{+}{+}$ may be expected if there are multiple very good systems, but controllability remains with $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$. Procedure 0-1 is less efficient and not more controllable.

5 Discussion and Conclusion

The choice of the selection procedure and its parameters can have a tremendous effect on the effort spent to select the best system, and the probability of making a correct decision. The new experimental setup (including random problem instances) and measurement displays (efficiency and target curves, as opposed to tables), proved useful for identifying strengths and weaknesses of both existing and new procedures.

For the SC and MDM configurations, the $\mathcal{LL}$ and $\mathcal{OCBA}_{\mathrm{LL}}$ allocations together with the $\mathrm{EOC}_{\mathrm{Bonf}}$ stopping rule were generally the most efficient. $\mathcal{KN}{+}{+}$ was also very efficient when $k = 2$, with similar variances and low PICS values, but was less efficient otherwise. Procedures that use an indifference zone parameter, δ^* ($\mathcal{KN}{+}{+}$ and $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ stopping rule) were very sensitive to the value of δ^* (i.e., may require a lot of sampling). No procedure was particularly controllable for the SC and MDM configurations.

An arbitrary configuration encountered in practice is not likely to be an SC or MDM configuration. With randomized configurations (RPI1 and RPI2), the $\text{PGS}_{\text{Slep},\delta^*}$ and EOC_{Bonf} stopping rules were reasonably controllable for the desired $\text{PGS}_{\text{IZ},\delta^*}$ and EOC_{IZ}, respectively. Procedures $\mathcal{LL}(\text{PGS}_{\text{Slep},\delta^*})$, $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$, $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$ and $\mathcal{KN}{+}{+}$ were the most efficient. Procedure $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$ uses a different allocation and stopping rule than is usual for the OCBA approach. Procedure $\mathcal{KN}{+}{+}$ tended to overdeliver for $\text{PGS}_{\text{IZ},\delta^*}$ in RPI experiments.

Strengths of $\mathcal{KN}{+}{+}$ include a low sensitivity to the number of first stage samples (n_0), and its natural ability to account for correlated output. It is the only method tested here with a proven asymptotic lower bound for $\text{PCS}_{\text{IZ}} \geq 1 - \alpha^* > 1/k$. The cost of preferring a lower bound for PCS is the potential for highly excessive sampling. Although we only tested moderate numbers of systems, $\mathcal{KN}{+}{+}$ seems to lose efficiency relative to the Bayesian procedures as the number of systems increases.

We recommend combining the Bayesian allocation procedures with an adaptive stopping rule to substantially improve efficiency. Independent of the stopping rule, the loss-based allocations $\mathcal{LL}$ and $\mathcal{OCBA}_{\text{LL}}$ are among the most efficient allocations. The Online Companion suggests theoretical reasons why $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ perform so similarly. The most efficient and controllable stopping rule depends on the desired goal (EOC_{Bonf} or $\text{PGS}_{\text{Slep},\delta^*}$). The strong efficiency is relatively robust to different configurations, and controllability is relatively robust for RPI. These procedures also allow for the incorporation of prior information about problem instances when that is available. Other strengths include the ability (i) to allocate samples in order to optimize the total CPU time even if the CPU times of each system are different (Chick and Inoue 2001a, and Appendix B), (ii) to run with a time constraint with the $\mathcal{S}$ stopping rule, and (iii) to run as a two-stage procedure if needed (e.g., the EOC_{Bonf} stopping rule can be adapted to two stages by finding a second-stage sampling budget that achieves a desired predicted EOC_{Bonf}). Weak points of those procedures are a dependency on the initial number of samples for each system, n_0 (we recommend $n_0 \geq 6$), and the potential for a small degradation in performance if there are many systems that are close to the best.

Procedures Equal and 0-1 are not recommended for general use.

We did not assess output from steady-state simulations. $\mathcal{KN}{+}{+}$ uses batching to estimate variances, a standard technique for the handling the autocorrelation from such simulations, but can take observations one-at-a-time. We do not see why batching might affect the different Bayesian procedures differently.

We therefore select $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$, with $\delta^* > 0$, for further work to integrate selection procedures into discrete optimization via simulation environments where PGS is of interest. The clear winners for applications where the simulation output represents economic value are $\mathcal{LL}(\text{EOC}_{\text{Bonf}})$ and $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$. Future goals are theory and practical developments to assess the use of CRN with those procedures in fully sequential environments, (ground work was laid by Fu et al. 2006 for the OCBA and by Chick and Inoue 2001b for the VIP), and the ability to further account for the economics of decisions made with simulation.

Acknowledgment. We thank Seong-hee Kim and Demet Batur for very helpful discussions about $\mathcal{KN}{+}{+}$. We thank three anonymous referees and the associate editor for constructive feedback.

REFERENCES

Bechhofer, R. E., T. J. Santner, and D. M. Goldsman (1995). *Design and Analysis for Statistical Selection, Screening, and Multiple Comparisons*. New York: John Wiley & Sons, Inc.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. Chichester, UK: Wiley.

Boesel, J., B. L. Nelson, and S.-H. Kim (2003). Using ranking and selection to 'clean up' after simulation optimization. *Operations Research 51*, 814–825.

Branke, J., S. E. Chick, and C. Schmidt (2005). New developments in ranking and selection, with an empirical comparison of the three main approaches. In M. Kuhl, N. Steiger, F. Armstrong, and J. Joines (Eds.), *Proc. 2005 Winter Simulation Conference*, Piscataway, NJ, pp. 708–717. IEEE, Inc.

Branke, J. and C. Schmidt (2004). Sequential sampling in noisy environments. In X. Yao et al. (Ed.), *Parallel Problem Solving from Nature*, Volume 3242 of *LNCS*, pp. 202–211. Springer.

Butler, J., D. J. Morrice, and P. W. Mullarkey (2001). A multiple attribute utility theory approach to ranking and selection. *Management Science 47*(6), 800–816.

Chen, C.-H. (1996). A lower bound for the correct subset-selection probability and its application to discrete event simulations. *IEEE Transactions on Automatic Control 41*(8), 1227–1231.

Chen, C.-H., J. Lin, E. Yücesan, and S. E. Chick (2000). Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems: Theory and Applications 10*(3), 251–270.

Chen, C.-H., E. Yücesan, L. Dai, and H. Chen (2006). Efficient computation of optimal budget allocation for discrete event simulation experiment. *IIE Transactions*, to appear.

Chen, E. J. and W. D. Kelton (2005). Sequential selection procedures: Using sample means to improve efficiency. *European Journal of Operational Research 166*, 133–153.

Chick, S. E. and K. Inoue (1998). Sequential allocation procedures that reduce risk for multiple comparisons. In D. J. Medeiros, E. J. Watson, M. Manivannan, and J. Carson (Eds.), *Proc. 1998 Winter Simulation Conference*, Piscataway, NJ, pp. 669–676. IEEE, Inc.

Chick, S. E. and K. Inoue (2001a). New two-stage and sequential procedures for selecting the best simulated system. *Operations Research 49*(5), 732–743.

Chick, S. E. and K. Inoue (2001b). New procedures for identifying the best simulated system using common random numbers. *Management Science 47*(8), 1133–1149.

Chick, S. E. and K. Inoue (2002). Corrigendum: New selection procedures. *Operations Research 50*(3), 566.

Chick, S. E. and Y. Wu (2005). Selection procedures with frequentist expected opportunity cost bounds. *Operations Research 53*(5), 867–878.

Dai, L. (1996). Convergence properties of ordinal comparison in the simulation of discrete event systems. *J. Optimization Theory and Applications 91*(2), 363–388.

de Groot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Fu, M. C., J.-Q. Hu, C.-H. Chen, and X. Xiong (2006). Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS Journal on Computing*, to appear.

Goldsman, D., S.-H. Kim, W. S. Marshall, and B. L. Nelson (2002). Ranking and selection for steady-state simulation: Procedures and perspectives. *INFORMS Journal on Computing 14*(1), 2–19.

Gupta, S. S. and K. J. Miescke (1994). Bayesian look ahead one stage sampling allocations for selecting the largest normal mean. *Statistical Papers 35*, 169–177.

He, D., S. E. Chick, and C.-H. Chen (2006). The opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Trans. Systems, Machines, Cybernetics C*, to appear.

Inoue, K., S. E. Chick, and C.-H. Chen (1999). An empirical evaluation of several methods to select the best system. *ACM TOMACS 9*(4), 381–407.

Kim, S.-H. and B. L. Nelson (2001). A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS 11*, 251–273.

Kim, S.-H. and B. L. Nelson (2006). Selecting the best system. In S. G. Henderson and B. L. Nelson (Eds.), *Handbook in Operations Research and Management Science: Simulation*. Elsevier.

Law, A. M. and W. D. Kelton (2000). *Simulation Modeling & Analysis* (3rd ed.). New York: McGraw-Hill.

Nelson, B. L. and S. Banerjee (2001). Selecting a good system: Procedures and inference. *IIE Transactions 33*(3), 149–166.

Nelson, B. L. and D. Goldsman (2001). Comparisons with a standard in simulation experiments. *Management Science 47*(3), 449–463.

Online Companion For:
Selecting a Selection Procedure

Jürgen Branke
Institute AIFB
University of Karlsruhe
Germany
branke@aifb.uni-karlsruhe.de

Stephen E. Chick
INSEAD
Technology Management Area
France
stephen.chick@insead.edu

Christian Schmidt
Institute AIFB
University of Karlsruhe
Germany
csc@aifb.uni-karlsruhe.de

This Online Companion to "Selecting a Selection Procedure" contains several technical appendices that present both empirical and analytical results.

Appendix A provides additional graphs that support the claims in the main paper. Appendix A.7 also provides an initial exploration of the distribution of the number of samples for different procedures, as an extension to the paper's focus on the expected number of samples as a measure of efficiency.

Appendix B provides a theoretical motivation to explain why $\mathcal{LL}$ and $\mathcal{OCBA}_{\text{LL}}$ perform similarly. Along the way, it shows how the $\mathcal{OCBA}_{\text{LL}}$ procedure can be extended to account for different sampling costs for each system (e.g., different CPU times), and how $\mathcal{OCBA}_{\text{LL}}$ might be run as a two-stage, rather than sequential, procedure, if that is desired. Those are properties of the original $\mathcal{LL}$ procedure.

Appendix C describes implementation issues, and describes computational techniques that can overcome numerical stability problems that arise in extreme applications of ranking and selection procedures. As it turns out, Procedures $\mathcal{KN}{+}{+}$, $\mathcal{LL}$, and 0-1 did not suffer from numerical stability issues in our experiments. $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\text{LL}}$ needed mild assistance when pushed to extreme levels of evidence for correct selection.

Appendix D provides additional information about the configurations that were tested for this paper.

A Additional Supporting Graphs

The paper presented a summary of the general conclusions from the study. This section contains a subset of additional results that explore the ideas further. It is not practical to display all results from the experiments, as we tested over 150 problem configurations defined by combinations of k, $\{$SC, MDM, RPI1, RPI2$\}$, and configuration parameters (δ, ρ for SC, MDM; η, b, c for RPI1, RPI2). Together with combinations of n_0, sampling allocations, and stopping rule parameters, well over 10^4 different combinations were run.

We developed a GUI to allow a graphical visualization and easy navigation through the results. That GUI was used to generate most of the figures in this paper.

A.1 Sample Variance Updates for Procedure $\mathcal{KN}{+}{+}$

The version of $\mathcal{KN}{+}{+}$ in §1.2 of the main paper updates the sample variances after each sample observed. Updating the sample variance means that asymptotic arguments are used to assess asymptotic PCS guarantees. The original Procedure $\mathcal{KN}$ estimates the variances only once after the initial sampling stage and uses them throughout the run, never updating. We also tested intermittent updating of the sample variances, by regularly refreshing the variance only once every several iterations of the main loop in Procedure $\mathcal{KN}{+}{+}$.

Intermittent updating of the sample variances requires a small change in the procedure. The exponent of $-2/(n-1)$ in the formula for η of Step 3b of Procedure $\mathcal{KN}{+}{+}$ should be replaced with $-2/(n'-1)$, where n' is the number of independent samples that are used to compute the sample variance. Similarly, assign $h^2 \leftarrow 2\eta(n'-1)$ in that same step.

In each of the handful of SC, MDM and RPI configurations that we tested, updating after every sample is at least as efficient as less frequent updates. Figures 18 and 19 illustrate that point. The number that corresponds to each line in the graph is the number of samples between each update of the sampling interval. The line associated with the value 0 describes the curve associated with never updating the sample variance after the initial estimate is made, based upon the first stage of sampling ($\mathcal{KN}$). Updating after every sample also resulted in the best performance with respect to target curves, up to sampling noise.

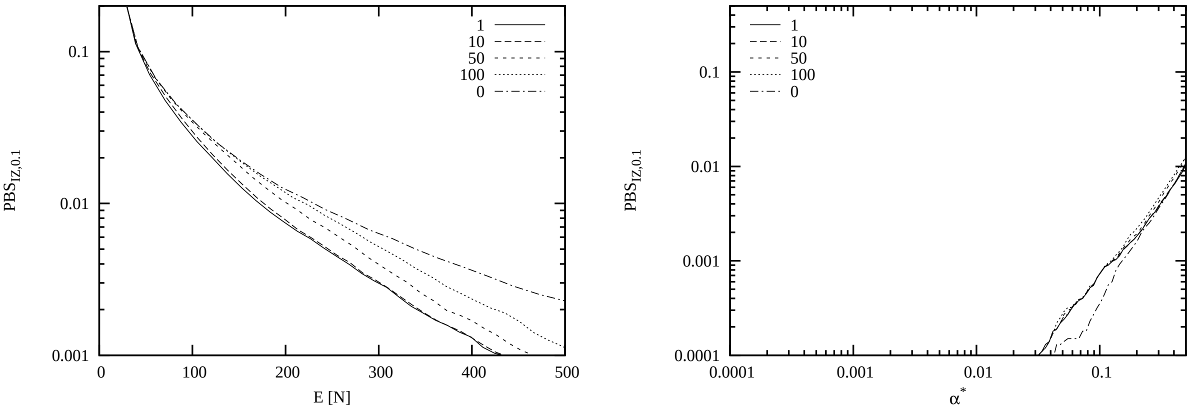


Figure 18: Influence of variance update intervals for $\mathcal{KN}{+}{+}_{0.1}$ (RPI1, $k=5$, $\eta=1$, $b=100$).

A.2 SC, $k=2$

The main paper indicated that the target curves for (α^*, $\mathrm{PCS}_{\mathrm{IZ}}$) and ($\alpha^*$, $\mathrm{PGS}_{\mathrm{IZ},\delta^*}$) for $\mathcal{KN}{+}{+}$ and $\mathrm{PGS}_{\mathrm{Slep},\delta^*}$ are sensitive to the choice of δ^*. Another way to assess that sensitivity for $\mathrm{PGS}_{\mathrm{IZ},\delta^*}$ with α^* is to fix α^*, and attempt to pick an δ^* such that the empirical $\mathrm{PCS}_{\mathrm{IZ}}$ matches the actually desired $1-\alpha^*$. While this is *not* the traditional way to set δ^*, this thought experiment will help explain how difficult it is to control a procedure to obtain exactly the desired $\mathrm{PCS}_{\mathrm{IZ}}$, rather than obtaining a lower bound.

Figure 20 shows the influence of δ^* on $\mathrm{PCS}_{\mathrm{IZ}}$ efficiency (the mean number of samples required to

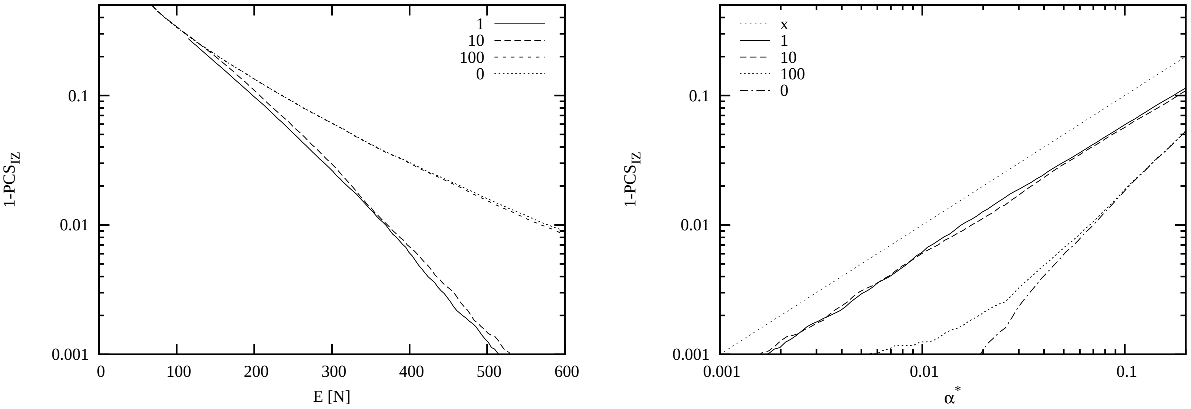


Figure 19: Influence of variance update intervals for $\mathcal{KN}{+}{+}_{0.5}$ (SC, $k=10$, $\delta=0.5$, $\rho=1$).

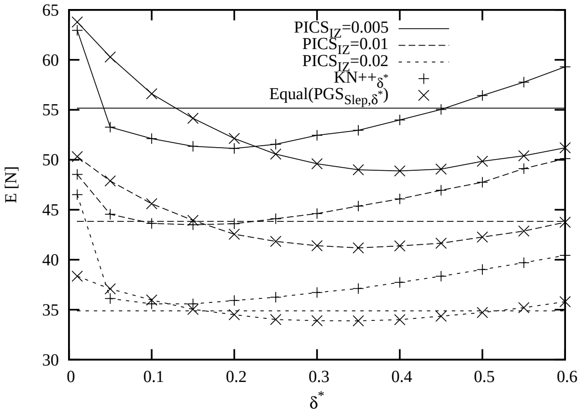


Figure 20: Influence of δ^* on mean number of samples to obtain a desired PICS$_{\text{IZ}}$, for $\mathcal{KN}{++}$ and Equal$(\text{PGS}_{\text{Slep},\delta^*})$ (SC, $k = 2$, $\delta = 0.5$, $\rho = 1$).

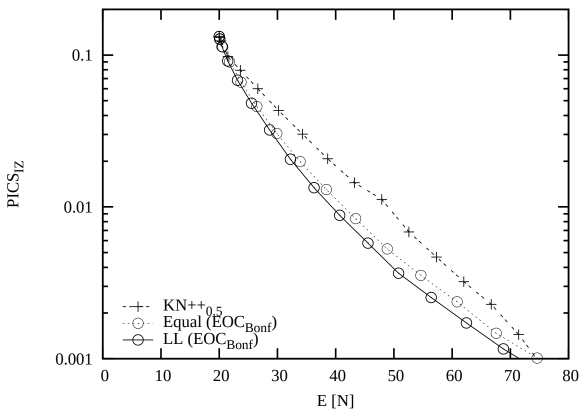


Figure 21: Different variances make Equal and $\mathcal{KN}{++}$ suboptimal with $k = 2$ (SC, $k = 2$, $\delta = 0.5$, $\rho = 0.354$, $\tau_0 = 10$).

obtain a specified level PICS$_{\text{IZ}}$). For small PICS$_{\text{IZ}}$, there may exist settings for δ^* so that $\mathcal{KN}{++}$ and Equal$(\text{PGS}_{\text{Slep},\delta^*})$ are more efficient than Equal(EOC$_{\text{Bonf}}$). To see this, note that when PICS$_{\text{IZ}} = 0.005$ and 0.01, the curves for $\mathcal{KN}{++}$ and Equal$(\text{PGS}_{\text{Slep},\delta^*})$ go below the horizontal lines. The horizontal lines show mean number of samples required by Equal(EOC$_{\text{Bonf}}$) to reach the corresponding PICS$_{\text{IZ}}$ level. The value of δ^* that is needed to obtain the minimal mean number of samples depends on the problem instance. Since the problem instance is unknown in practice, it is not clear how to set δ^* in general.

The main paper primarily focused on the case of equal variances when $k = 2$. It is not optimal to sample from each system equally often when $k = 2$ and the variances of each system differ. The Equal and $\mathcal{KN}{++}$ allocations both sample equally often when $k = 2$. The Bayesian allocations do not need to sample equally often, and therefore can be more efficient than procedures that sample equally often when $k = 2$. Figure 21 illustrates that point.

A.3 SC, $k > 2$

Figure 22 (left panel) illustrates the observation that the advantage of adaptive Bayesian procedures, relative to Equal, increases with k. The qualitative nature of the graph does not change as δ and ρ are individually varied from the values used for the plot. For all other procedures with the new stopping rules, there is a tendency to overdeliver as k increases, but $\mathcal{OCBA}_{\text{LL}}$ is more sensitive than Equal (right panel). The tendency to overdeliver with increasing k might be attributed to the slack introduced by Slepian's and Bonferroni's inequalities.

Similar to Equal, $\mathcal{KN}{++}$ also loses efficiency relative to $\mathcal{OCBA}$(EOC$_{\text{Bonf}}$) as the number of systems k increases (Figure 23).

Figure 24 shows the importance of a sufficiently large n_0 for the case $k > 2$, too. The right panel indicates that increasing n_0 increases the tendency to overdeliver. While $\mathcal{LL}$(EOC$_{\text{Bonf}}$) is slightly closer to the target than $\mathcal{OCBA}_{\text{LL}}$(EOC$_{\text{Bonf}}$), it is slightly less efficient. $\mathcal{KN}{++}$ is insensitive to n_0 (this was demonstrated in Figure 5 for SC with $k = 2$, and is demonstrated for another configuration in Figure 31 below).

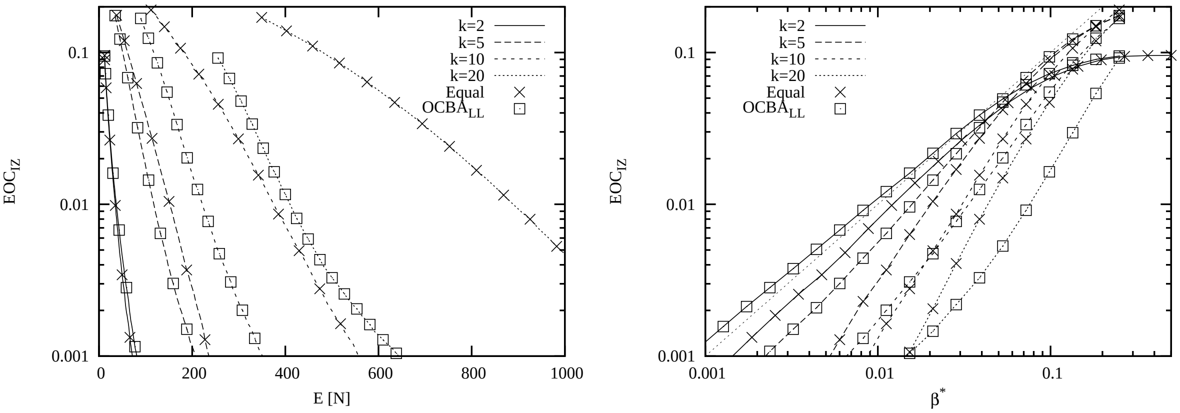


Figure 22: Influence of the number of systems k on efficiency (right panel) and target (left panel). Equal and $\mathcal{OCBA}$ allocation are used in combination with EOC$_{\text{Bonf}}$ stopping rule (SC, $\delta = 0.5$, $\rho = 1$).

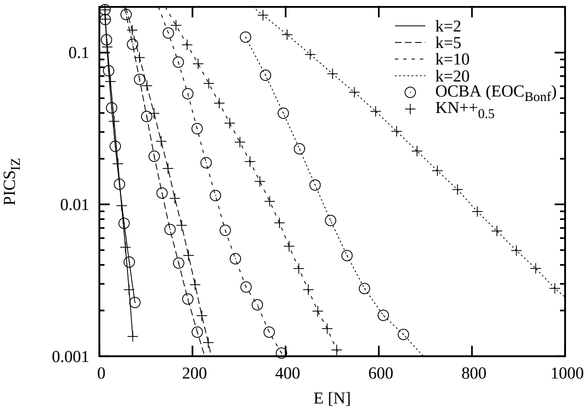


Figure 23: Efficiency of $\mathcal{KN}$++ and $\mathcal{OCBA}$(EOC$_{\text{Bonf}}$), as function of the number of systems k (SC, $\delta = 0.5$, $\rho = 1$).

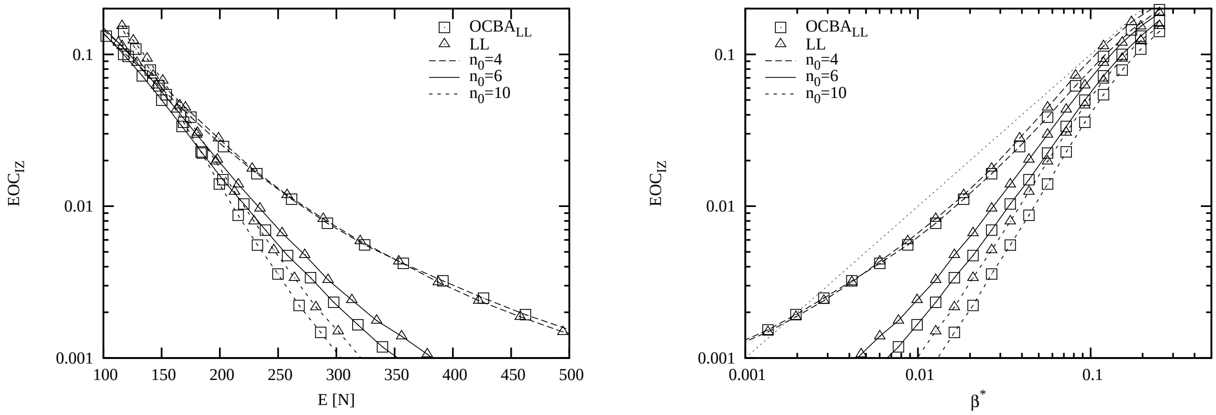


Figure 24: Effect of various n_0, with EOC$_{\text{Bonf}}$ stopping rule on *efficiency* (left) and *target* (right) (SC, $k = 10$, $\delta = 0.5$, $\rho = 1$).

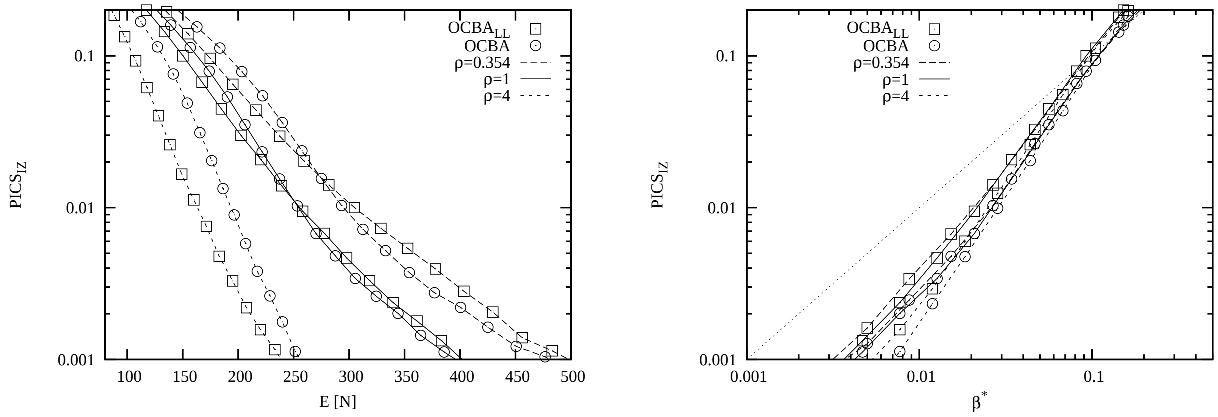


Figure 25: Influence of variance ratio ρ on $\mathcal{OCBA}$ and $\mathcal{OCBA}_{LL}$ (SC, $k = 10$, $\delta = 0.5$, EOC$_{\mathrm{Bonf}}$).

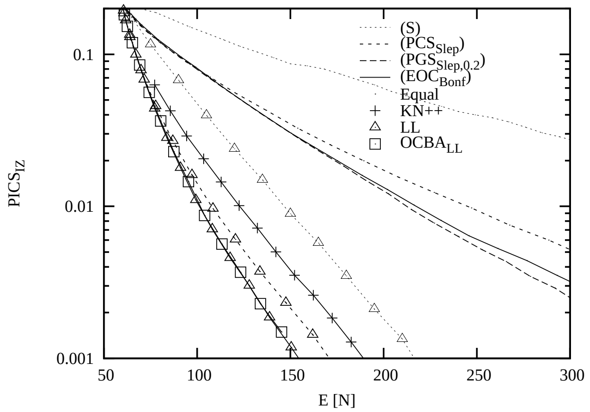


Figure 26: Different stopping rules (line types) and allocation procedures (symbols) (MDM, $k = 10$, $\delta = 0.5$, $\rho = 1$).

A larger variance for the best system relative to the other systems (larger ρ) makes correct selections easier (Figure 25, left panel). Different allocation functions respond differently to ρ, as illustrated for $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\mathrm{LL}}$, representatives for the PCS-based and EOC-based allocations. With $\rho = 4$, the EOC-based allocation is clearly superior, but the advantage diminishes for low PICS as ρ decreases, and the PCS-based allocation partially outperforms the EOC-based allocation. Larger ρ overdeliver slightly more on the target plots, and $\mathcal{OCBA}$ overdelivers slightly more than $\mathcal{OCBA}_{\mathrm{LL}}$ (right panel).

A.4 MDM, $k > 2$

First, recall that $\mathcal{LL}$ and $\mathcal{OCBA}_{\mathrm{LL}}$ performed virtually indistinguishably, except for sampling error, for all MDM settings that we examined. Thus we only show one of these two procedures in a given graph.

Figure 26 is like the efficiency plot in Figure 8 of the main paper, except that $k = 10$ instead of $k = 100$ systems are analyzed by the procedures. As would be expected, the number of samples required for $k = 100$ is much higher than for $k = 10$. The relative ordering of the procedures is the same, and the Equal allocation suffers by far more than the others as the number of systems is increased.

Figure 27 shows PBS$_{\mathrm{IZ},\delta^*}$ efficiency and target performance for different δ^*. The parameter δ^* for

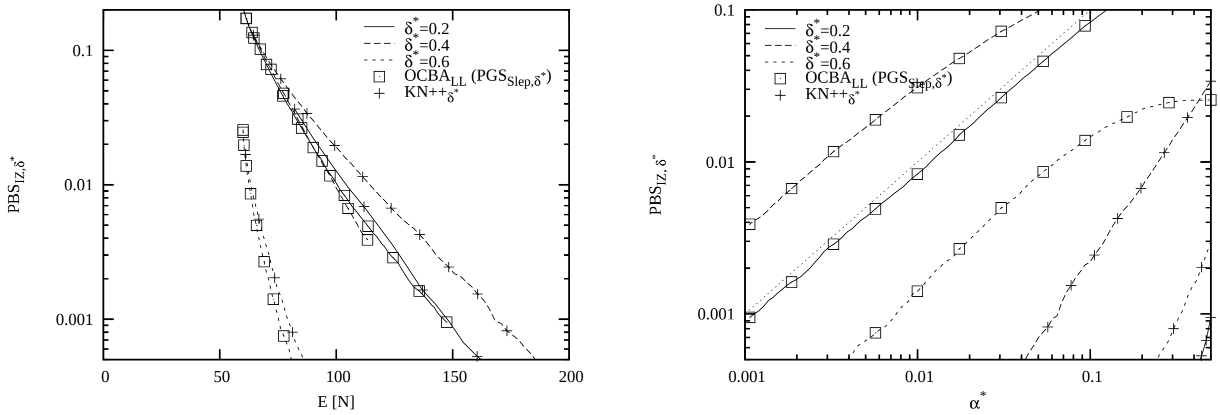


Figure 27: $\text{PBS}_{\text{IZ},\delta^*}$ efficiency and target for $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ and $\mathcal{KN}$++ for different settings of δ^* (MDM, $k = 10$, $\delta = 0.5$, $\rho = 1$). In this configuration $\text{PBS}_{\text{IZ},0.2} = \text{PBS}_{\text{IZ},0.4} = \text{PCS}_{\text{IZ}}$.

$\mathcal{KN}$++ and $\text{PGS}_{\text{Slep},\delta^*}$ stopping rule have thereby been set to the indifference zone used for measuring performance. For MDM with $\delta = 0.5$, an indifference zone of $\delta^* = 0.2$ or $\delta^* = 0.4$ are equivalent for $\text{PBS}_{\text{IZ},\delta^*}$ efficiency, as only the best system is considered to be a good selection. For $\delta^* = 0.6$, the two best systems are considered good. On the efficiency plot (left panel), it can be seen that the problem with $\text{PBS}_{\text{IZ},0.6}$ is significantly easier. The efficiency curves for $\delta^* = 0.2$ and 0.4 are very similar for $\mathcal{OCBA}_{\text{LL}}$, while $\mathcal{KN}$++ loses efficiency for $\delta^* = 0.4$. On the target plot (right panel), the target performance of both procedures is affected by δ^*. Overall, the $\text{PGS}_{\text{Slep},\delta^*}$ stopping rule is closer to the target than $\mathcal{KN}$++, which very much overdelivers in each case (the curve for $\delta^* = 0.2$ is almost outside the plot). On the other hand, $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ consistently underdelivers for $\delta^* = 0.4$.

A.5 RPI1

Figure 28 compares three selection procedures with adaptive stopping rules, $\text{Equal}(\text{PGS}_{\text{Slep},\delta^*})$, $\mathcal{KN}$++, and $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$. As is typical for the RPI1 problems tested, $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$ outperforms $\mathcal{KN}$++ for efficiency (left panel) and controllability (right panel). When moving from $b = 100$ (very similar variances for each system) to $b = 2.5$ (very different variances), the efficiency of $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$ improves and the efficiency of $\mathcal{KN}$++ is basically not affected.

Figure 29 further illustrates this comparison, and is similar to Figure 14 of the main paper except that $k = 5$ instead of $k = 50$. Figure 29 depicts the effect of different δ^* and α^* on the number of samples required by $\mathcal{KN}$++ and $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$, and the ability to deliver the desired PGS (the δ^* parameter of the procedure is also used to measure $\text{PGS}_{\text{IZ},\delta^*}$). The left panel shows that $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ follows the target better than $\mathcal{KN}$++, although $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ slightly underdelivers for low values of α^*. Again, the right panel shows that the penalty for conservativeness is a significantly higher sampling effort for a given desired α^*.

One question is whether the δ^* of the selection procedure can be selected in a nontraditional way to achieve a given desired performance for the probability of a good selection. That is, one might ask if it is possible to choose a value of δ^* for a procedure's allocation and stopping rule in a way that controls the desired empirical performance for $\text{PGS}_{\delta^{**}}$, for some $\delta^{**} \neq \delta^*$. Here we explore a response to that

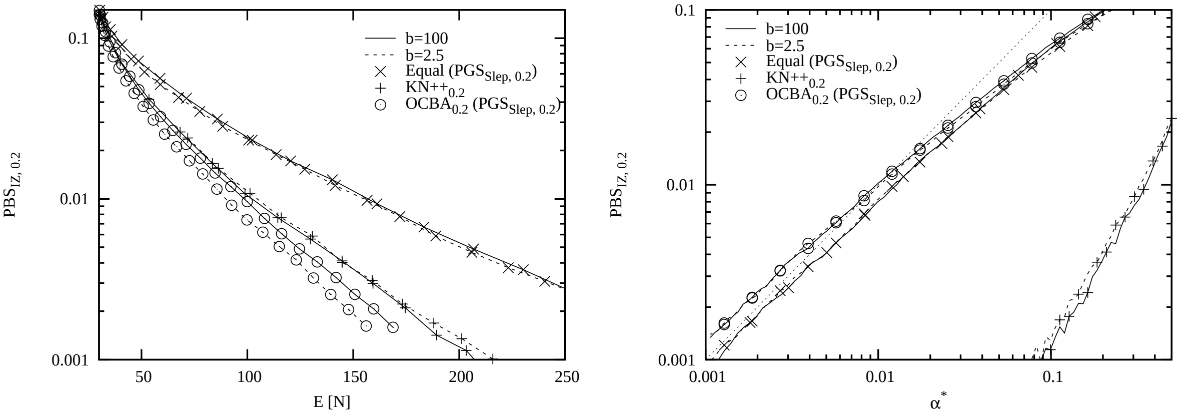


Figure 28: Efficiency and target for PCS-based procedures (RPI1, $k = 5$, $\eta = 1$, $b \in \{2.5, 100\}$).

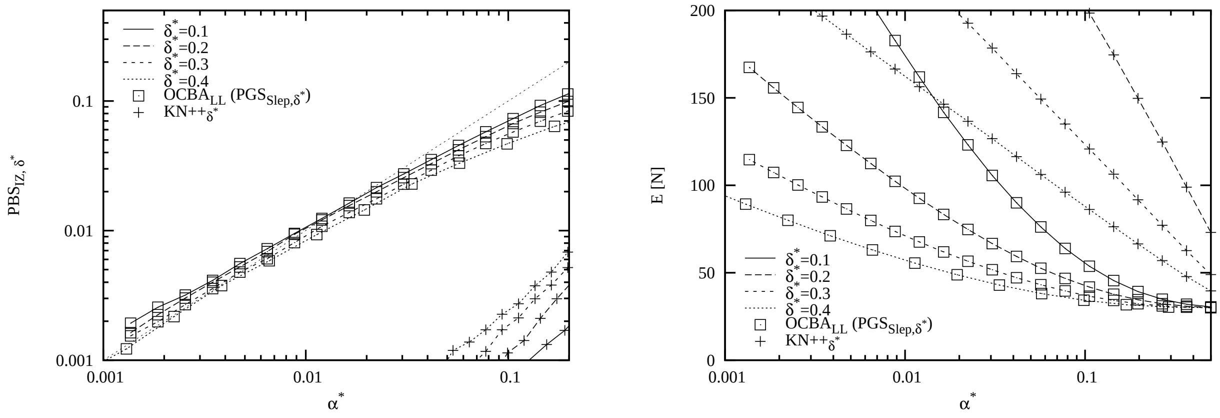


Figure 29: Comparison of $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ and $\mathcal{KN}{+}{+}$ for $\text{PBS}_{\text{IZ},\delta^*}$ target (left panel) and number of samples (right panel) for different δ^* and α^* (RPI1, $k = 5$, $\eta = 1$, $b = 100$).

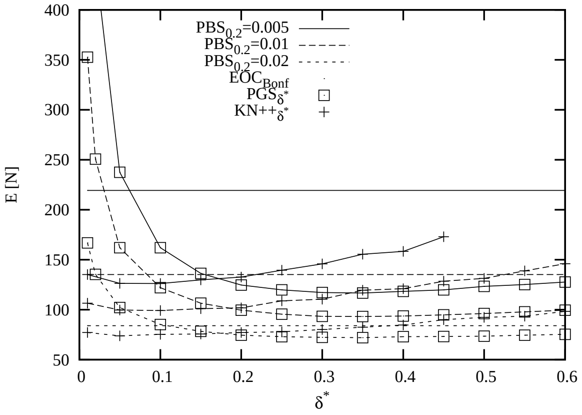


Figure 30: Influence of δ^* on the required number of samples to obtain a desired $\text{PBS}_{\text{IZ},0.2} = \alpha^*$ for $\mathcal{KN}{++}$ and $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$ (RPI1, $k = 5$, $\eta = 1, b = 100$).

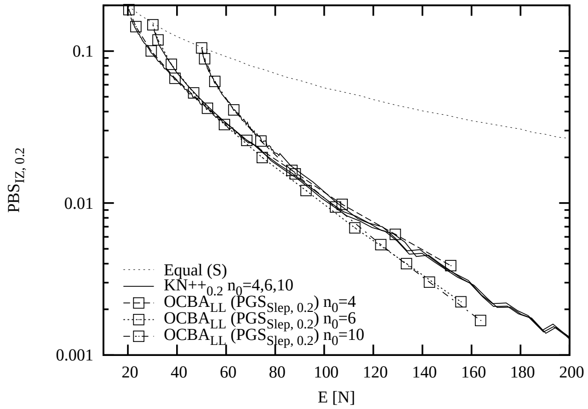


Figure 31: Efficiency of $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{KN}{++}$ depends on n_0 (RPI1, $k = 5, \eta = 1, b = 100$).

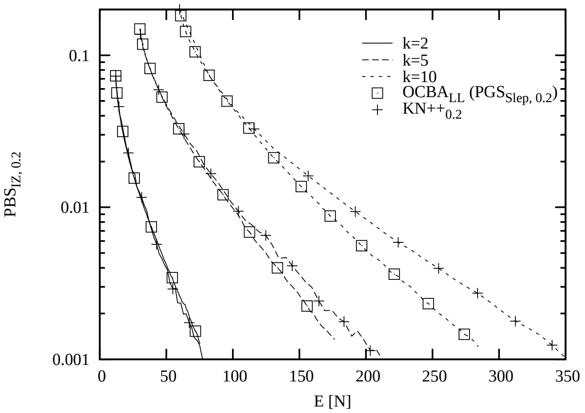


Figure 32: Efficiency depends on the number of systems k (RPI1, $\eta = 1, b = 100$).

question with empirical results. Figure 30 illustrates the influence of δ^* on the $\text{PGS}_{\text{IZ},0.2}$ efficiency of $\mathcal{KN}{++}$ and $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$. For reference, the horizontal lines show the number of samples required by $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$. For RPI, setting the procedure's parameter δ^* to the δ^* as specified in the efficiency goal (0.2) yields reasonable, though not optimal, efficiency for both procedures. The target performance is good for $\mathcal{OCBA}_{\text{LL}}(\text{PGS}_{\text{Slep},\delta^*})$, while $\mathcal{KN}{++}$ significantly overdelivers (Figure 28, right panel).

Some other observations from SC and MDM also carry over to RPI: The Bayesian procedures are generally very sensitive to n_0, while $\mathcal{KN}{++}$ is not (Figure 31), and $\mathcal{KN}{++}$ becomes less efficient relative to the Bayesian procedures as the number of systems k increases (Figure 32).

Figure 33 shows that the benefit of including prior information in the VIP and OCBA procedures is more or less independent of b and η for the values tested.

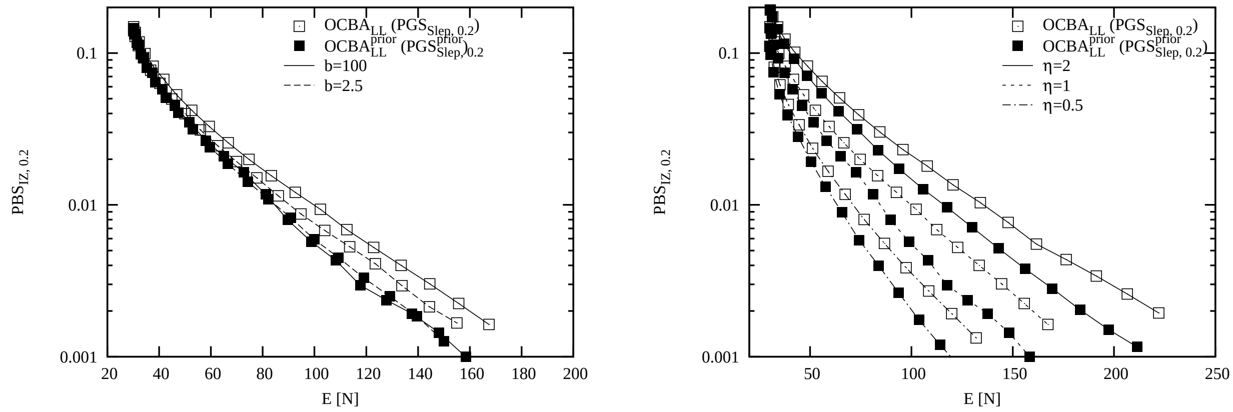


Figure 33: Benefit of providing prior information depending on problem configuration parameters b, η (RPI1, $k = 5$, $b = 100$, unless specified otherwise).

A.6 RPI2

§4 states that most observations made for RPI1 carry over to RPI2 even in the case of $a = 1$, i.e. many good systems. Some evidence for this claim is given here. Figure 34 shows that the main conclusions about allocation rules also hold for RPI2. The $\mathcal{LL}$, $\mathcal{OCBA}_{\rm LL}$ and $\mathcal{OCBA}_{\delta^*}$ allocations are more or less equally efficient. Procedures 0-1 and Equal are clearly less efficient.

The relative ordering of the stopping rules with respect to EOC$_{\rm IZ}$ efficiency in combination with the Equal allocation remains the same: PGS$_{{\rm Slep},\delta^*}$ is more efficient than EOC$_{\rm Bonf}$ which is more efficient than $\mathcal{S}$ stopping rule (Figure 35). Figure 36 compares target plots for the negative exponential (RPI2, $a = 1$), Gaussian (RPI1) and positive exponential (RPI2, $a = 0$) distribution of the means, in the order of decreasing number of good systems. If the sampling distribution for the means matches the prior distribution used for a Bayesian procedure, the target is matched closely. Modifying the distribution towards more good systems (negative exponential) or fewer good systems (positive exponential) leads to over- and underdelivery, respectively.

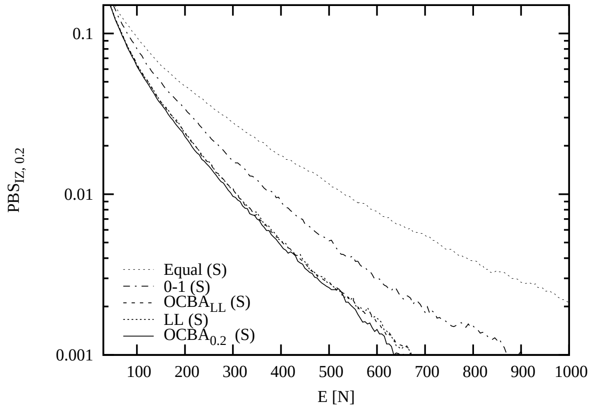


Figure 34: PBS$_{{\rm IZ},\delta^*}$ efficiency of allocations with $\mathcal{S}$ stopping rule (RPI2, $k = 5, a = 1, \eta = 1, b = 100$).

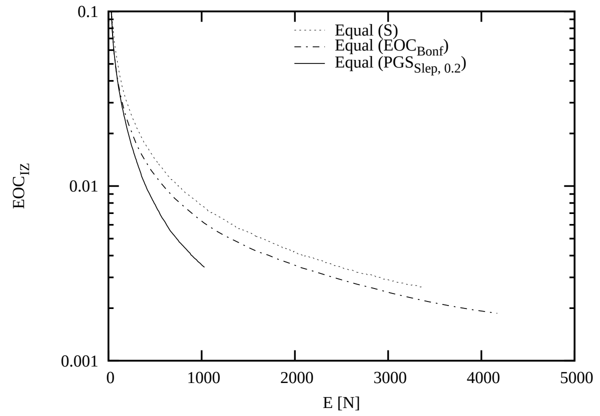


Figure 35: Efficiency for Equal allocation and different stopping rules w.r.t. EOC_{IZ} (RPI2, $k = 5$, $a = 1$, $\eta = 1$, $b = 100$).

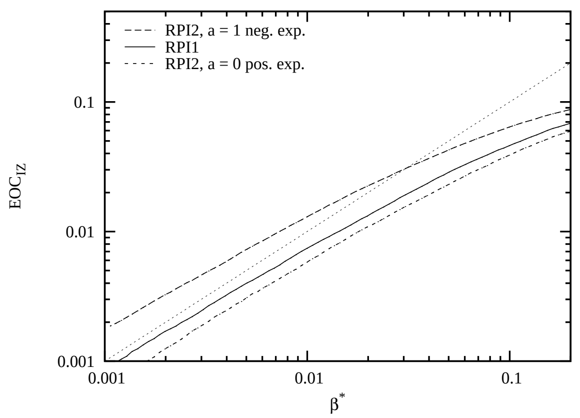


Figure 36: Influence of the sampling distributions of configurations on target (RPI1, RPI2, $k = 5, \eta = 1, b = 100$, $\text{Equal}(\text{EOC}_{\text{Bonf}})$).

A.7 Distribution of the Number of Samples

The main paper describes how the mean number of samples of each procedure can vary as a function of the allocation rule, stopping rule, and the parameters of each individual procedure. This subsection looks at the distribution of the number of samples for some of the procedures for some representative configurations. We did not do an exhaustive analysis over all configurations, parameter values, allocations, and so forth for this subsection, but the following graphs give some initial ideas of how the procedures perform in distribution, rather than on average.

We first note that with the $\mathcal{S}$ stopping rule, one can completely control the number of samples. The distribution of the number of samples equals the mean number of samples for that stopping rule, independent of the configuration and of the allocation rule.

We now turn to flexible stopping rules. A direct comparison of the distribution of the number of samples is not quite obvious, since picking the same parameters for each procedure leads to a different mean number of samples and a different PCS_{IZ} or EOC_{IZ}. We therefore attempt to display the distributions by picking the parameters of each procedure separately, in order to obtain a similar empirical figure of merit, as well as a more direct comparison of the procedures as a function of their parameters.

Figure 37 shows the distribution of the number of samples that several procedures require, if the stopping rule parameters are chosen so that each procedure stops, on average, after approximately 400 samples. In Figure 38, the stopping rules for each procedure were chosen to achieve $\text{PICS}_{\text{IZ}} = 0.005$. Table 1 provides the specific parameter values and the estimates of the figures of merit for each procedure. The distributions are rather more skewed for $\text{Equal}(\text{EOC}_{\text{Bonf}})$ and $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$ than for $\mathcal{KN}$++, when parameters are chosen to achieve the same mean number of samples.

The distribution of the number of samples for different stopping values are shown in Figure 39 and Figure 40. While the main paper indicates that the EOC_{Bonf} stopping rule with $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ allocations tend to be more efficient than $\mathcal{KN}$++, these graphs indicate that the distribution of the number of samples that are required by $\mathcal{KN}$++ has a smaller 'tail' than that for $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$. That is, there is a high

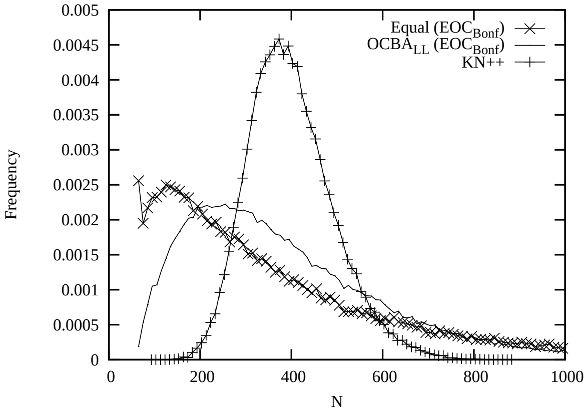


Figure 37: Distribution of the number of samples for Equal$(\text{EOC}_{\text{Bonf}} = 0.0158)$, $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}} = 0.00107)$ and $\mathcal{KN}{+}{+}_{\delta^*=0.2}(\alpha^* = 0.191)$ (MDM, $k = 10$, $\delta = 0.25$, $\rho = 1$, $\text{E}[N] \approx 400$).

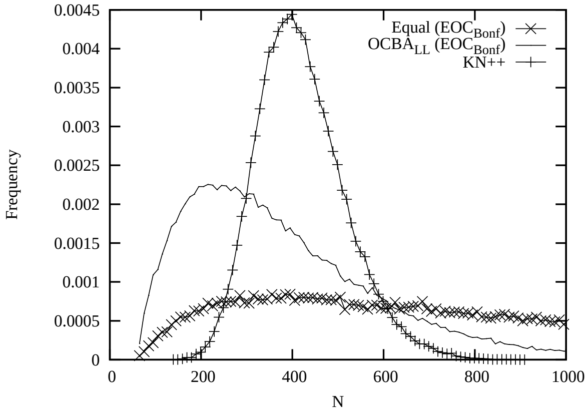


Figure 38: Distribution of the number of samples for Equal$(\text{EOC}_{\text{Bonf}} = 0.00174)$, $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}} = 0.00115)$ and $\mathcal{KN}{+}{+}_{\delta^*=0.2}(\alpha* = 0.166)$ (MDM, $k = 10$, $\delta = 0.25$, $\rho = 1$, PICS ≈ 0.005).

Procedure: allocation (stopping rule)	$\text{E}[N]$	Stddev$[N]$	PICS$_{\text{IZ}}$	EOC$_{\text{IZ}}$
Equal$(\text{EOC}_{\text{Bonf}} \leq 0.0158489)$	**393.873**	332.001	0.06981	0.0183
Equal$(\text{EOC}_{\text{Bonf}} \leq 0.0017378)$	1064.7	849.368	**0.00525**	0.0013275
$\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}} \leq 0.00114815)$	384.794	223.046	**0.00512**	0.0013125
$\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}} \leq 0.00107152)$	**391.677**	226.293	0.00477	0.0012225
$\mathcal{KN}{+}{+}_{\delta^*=0.2}(\alpha^* = 0.190546)$	**399.917**	93.2954	0.00643	0.00161
$\mathcal{KN}{+}{+}_{\delta^*=0.2}(\alpha^* = 0.165959)$	420.515	96.2242	**0.00505**	0.001265

Table 1: Parameters and estimated figures of merit for MDM comparisons in Figures 37–38.

probability that $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ will require fewer samples than $\mathcal{KN}{+}{+}$, but there is a small probability that $\mathcal{OCBA}_{\text{LL}}$ and $\mathcal{LL}$ will require many more samples than $\mathcal{KN}{+}{+}$.

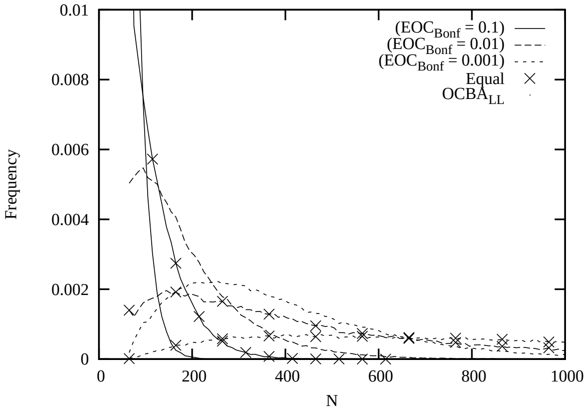


Figure 39: Distribution of the number of samples for Equal$(\text{EOC}_{\text{Bonf}})$ and $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$ (MDM, $k = 10$, $\delta = 0.25$, $\rho = 1$).

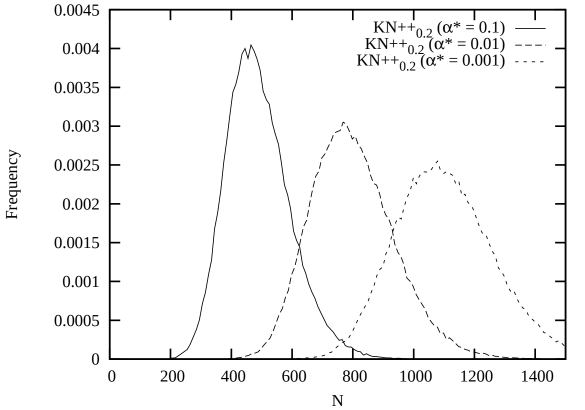


Figure 40: Distribution of the number of samples for $\mathcal{KN}{+}{+}$ (MDM, $k = 10$, $\delta = 0.25$, $\rho = 1$).

Figure 41 and Figure 42 are analogous to Figure 37 and Figure 38, except that they apply to an RPI1 configuration rather than to an MDM configuration. They show that when the parameters of each procedure

Procedure: allocation (stopping rule)	E$[N]$	Stddev$[N]$	PGS$_{\text{IZ},0.2}$
$\mathcal{OCBA}_{\delta^*=0.2}(1-\text{PGS}_{\text{Slep},0.2} \leq 0.00524807)$	201.778	166.606	**0.00515**
$\mathcal{OCBA}_{\delta^*=0.2}(1-\text{PGS}_{\text{Slep},0.2} \leq 0.00199526)$	**250.672**	214.615	0.00218
$\mathcal{KN}{+}{+}_{\delta^*=0.2}(\alpha^* = 0.40738)$	234.541	107.828	**0.0052**
$\mathcal{KN}{+}{+}_{\delta^*=0.2}(\alpha^* = 0.354813)$	**254.658**	118.029	0.00397

Table 2: Parameters and estimated figures of merit for RPI comparisons in Figures 41-42.

are selected separately in order to obtain a similar mean number of samples, or a similar empirical probability of bad selection, then the Bayesian procedures again have a larger skew. For all procedures, the distributions for RPI are more skewed than for MDM, presumably due to some very hard random problem instances. Table 2 provides the specific parameter values and the estimates of the figures of merit for each procedure for these graphs.

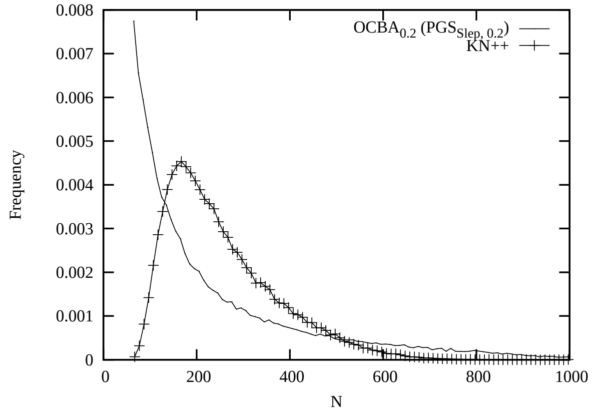


Figure 41: Distribution of the number of samples for $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$ and $\mathcal{KN}{+}{+}$ (RPI1, $k = 10$, $\eta = 1$, $b = 100$, E$[N] \approx 250$).

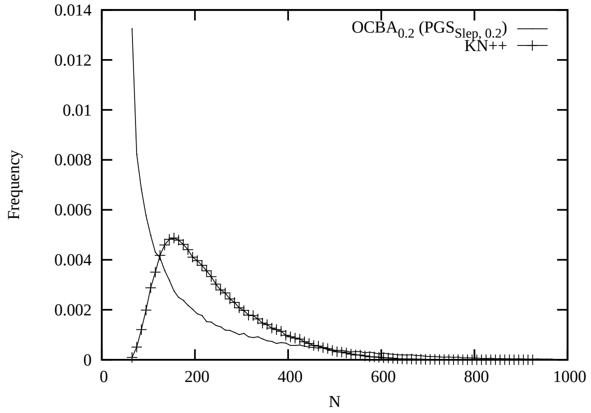


Figure 42: Distribution of the number of samples for $\mathcal{OCBA}_{\text{LL}}(\text{EOC}_{\text{Bonf}})$ and $\mathcal{KN}{+}{+}$ (RPI1, $k = 10$, $\eta = 1$, $b = 100$, PGS$_{\text{IZ},0.2} \approx 0.005$).

If, on the other hand, the parameters for the Bayesian procedures and $\mathcal{KN}{+}{+}$ are chosen similarly, as for $\mathcal{OCBA}_{\delta^*}(\text{PGS}_{\text{Slep},\delta^*})$ and $\mathcal{KN}{+}{+}$, the mean number of samples for the Bayesian procedures is significantly smaller than that of $\mathcal{KN}{+}{+}$, and the 'tail' of the distribution of the Bayesian procedures is not an issue, see for example Figure 43 and Figure 44.

In summary of the limited number of experiments to examine the distribution of samples, and not only the mean number of samples, the Bayesian procedures appear to have a larger skew than $\mathcal{KN}{+}{+}$, and all procedures have a larger skew for RPI than for MDM configurations.

The larger skew of the Bayesian procedures means that these procedures would have performed relatively even better if we had chosen the median instead of the mean number of replications to measure efficiency. On the other hand, they can require a much larger number of replications than $\mathcal{KN}{+}{+}$ in the worst case. However, at least for RPI configurations, this does not seem to be an issue, as the Bayesian procedures require significantly less samples for the equivalent parameter settings, an effect that hides the effect of skew.

These preliminary observations suggest potential value for a combined stopping rule for the Bayesian

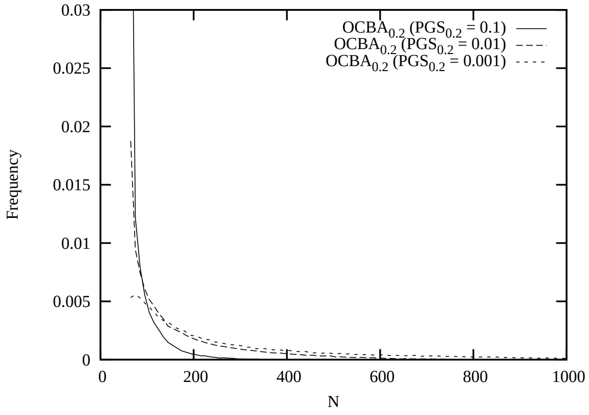


Figure 43: Distribution of the number of samples for $\mathcal{OCBA}_{0.2}(\text{PGS}_{\text{Slep},0.2})$ (RPI1, $k = 10$, $\eta = 1$, $b = 100$).

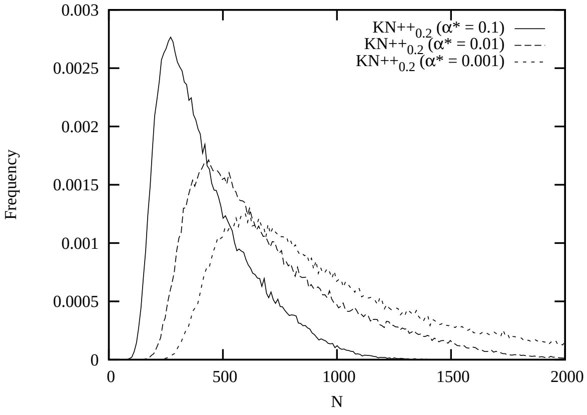


Figure 44: Distribution of the number of samples for $\mathcal{KN}{+}{+}$ (RPI1, $k = 10$, $\eta = 1$, $b = 100$).

procedures, which continue to allocate samples until either a flexible stopping rule is satisfied, or a budget limitation is reached. That rule would be easy to implement in practice (keep running until the evidence looks very strong or an analysis deadline is reached). A full analysis of that idea is reserved for future work.

B Asymptotic Relationship Between $\mathcal{LL}$ And $\mathcal{OCBA}_{\mathrm{LL}}$

The $\mathcal{LL}$ allocation is derived to maximize the value of information, relative to EOC$_{\mathrm{Bonf}}$, for asymptotically *large* numbers of additional replications per stage. This section shows that a continuous generalization of $\mathcal{OCBA}_{\mathrm{LL}}$ shares similar properties for asymptotically *small* numbers of additional replications. This property gives theoretical motivation why those two allocations empirically perform so similarly.

The $\mathcal{OCBA}_{\mathrm{LL}}$ presented above can be considered to be a discretized version of the following continuous optimization formulation which generalizes $\tilde{\lambda}_{ij}$ from (9) to the form (11) that allows every system to have a different number of additional replications, $\tau_{(i)}$, with the $\tau_{(i)}$ treated as real-valued (and not constrained to be nonnegative),

$$\begin{aligned} \text{EEOCS} &= \sum_{j:(j)\neq(k)} \tilde{\lambda}_{jk}^{-1/2}\Psi_{\tilde{\nu}_{(j)(k)}}\left[\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)}\right] \\ \tilde{\lambda}_{jk}^{-1} &= \frac{\hat{\sigma}^2_{(k)}}{n_{(k)}+\tau_{(k)}} + \frac{\hat{\sigma}^2_{(j)}}{n_{(j)}+\tau_{(j)}}. \end{aligned} \tag{11}$$

Suppose that $\tau_{(j)}$ replications of system (j) cost $c_{(j)}$ per replication. The optimal number of replications for each system, subject to a budget constraint $\tau = \sum_{i=1}^{k} c_i\tau_i$ can be determined with Lagrange multipliers. Let θ be the Lagrange multiplier associated with the sampling budget constraint, and let j be such that $(j) \neq (k)$. Recall $\Psi_\nu[s] = \frac{\nu+s^2}{\nu-1}\phi_\nu(s) - s(1-\Phi_\nu(s))$, and note that $\partial\Psi_\nu[s]/\partial s = \Phi_\nu(s) - 1$ and $\partial\tilde{\lambda}_{jk}/\partial\tau_{(j)} = \tilde{\lambda}_{jk}^2\hat{\sigma}^2_{(j)}/(n_{(j)}+\tau_{(j)})^2$. First order optimality conditions imply:

$$\begin{aligned} \theta c_{(j)} &= \frac{\partial\tilde{\lambda}_{jk}^{-1/2}}{\partial\tau_{(j)}}\Psi_{\tilde{\nu}_{(j)(k)}}[\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)}] + \tilde{\lambda}_{jk}^{-1/2}\frac{\partial\Psi_{\tilde{\nu}_{(j)(k)}}[\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)}]}{\partial\tau_{(j)}} \qquad (12) \\ &= -\frac{\tilde{\lambda}_{jk}^{1/2}}{2}\frac{\hat{\sigma}^2_{(j)}}{(n_{(j)}+\tau_{(j)})^2}\Psi_{\tilde{\nu}_{(j)(k)}}[\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)}] + \frac{\tilde{\lambda}_{jk}}{2}[\Phi_{\tilde{\nu}_{(j)(k)}}(\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)})-1]\frac{d_{(j)(k)}\hat{\sigma}^2_{(j)}}{(n_{(j)}+\tau_{(j)})^2} \\ &= -\frac{1}{2}\frac{\tilde{\nu}_{(j)(k)}+\tilde{\lambda}_{jk}d^2_{(j)(k)}}{\tilde{\nu}_{(j)(k)}-1}\phi_{\tilde{\nu}_{(j)(k)}}(\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)})\tilde{\lambda}_{jk}^{1/2}\frac{\hat{\sigma}^2_{(j)}}{(n_{(j)}+\tau_{(j)})^2} \\ &= -\frac{\hat{\sigma}^2_{(j)}\tilde{\zeta}_{(j)}}{2(n_{(j)}+\tau_{(j)})^2} \end{aligned}$$

where $\tilde{\zeta}_{(j)} = \frac{\tilde{\nu}_{(j)(k)}+\tilde{\lambda}_{jk}d^2_{(j)(k)}}{\tilde{\nu}_{(j)(k)}-1}\phi_{\tilde{\nu}_{(j)(k)}}(\tilde{\lambda}_{jk}^{1/2}d_{(j)(k)})\tilde{\lambda}_{jk}^{1/2}$ does not depend on any τ_i. Note that the optimal $\mathcal{OCBA}_{\mathrm{LL}}$ allocation with budget constraint seeks to preserve the following ratio between the total number of replications of systems (ℓ) and $(j) \neq (k)$:

$$\frac{n_{(\ell)}+\tau_{(\ell)}}{n_{(j)}+\tau_{(j)}} = \left(\frac{\hat{\sigma}^2_{(\ell)}\tilde{\zeta}_{(\ell)}/c_{(\ell)}}{\hat{\sigma}^2_{(j)}\tilde{\zeta}_{(j)}/c_{(j)}}\right)^{1/2}. \tag{13}$$

A similar analysis holds for system (k), with $\tilde{\zeta}_{(k)} = \sum_{j:(j)\neq(k)}\tilde{\zeta}_{(j)}$, and $\theta c_{(k)} = -\hat{\sigma}^2_{(k)}\tilde{\zeta}_{(k)}/2(n_{(k)}+\tau_{(k)})^2$.

The $\tilde{\zeta}_{(j)}$ depend upon the $n_{(\ell)}$ and $\tau_{(\ell)}$, but there is a 'fixed point' that preserves the fraction of replications allocated to each system from one stage to the next. Namely, if each $\tau_{(\ell)} \to 0$ as $\tau \to 0$, then $\tilde{\lambda}_{jk} \to \lambda_{jk}$,

$\tilde{\nu}_{(j)(k)} \rightarrow \nu_{(j)(k)}$, and $\tilde{\zeta}_{(j)} \rightarrow \zeta_{(j)}$ as $\tau \rightarrow 0$, where $\zeta_j = \frac{\nu_{(j)(k)}+\lambda_{jk}d^2_{(j)(k)}}{\nu_{(j)(k)}-1}\phi_{\nu_{(j)(k)}}(\lambda_{jk}^{1/2}d_{(j)(k)})\lambda_{jk}^{1/2}$. That means that *if* $\frac{n_{(\ell)}}{n_{(j)}} = \left(\frac{\hat{\sigma}^2_{(\ell)}\zeta_{(\ell)}/c_{(\ell)}}{\hat{\sigma}^2_{(j)}\zeta_{(j)}/c_{(j)}}\right)^{1/2}$ and each $\tau_{(\ell)} \rightarrow 0$ as $\tau \rightarrow 0$, *then* an infinitesimally small number of additional replications does not change that ratio,

$$\lim_{\tau\rightarrow 0} \frac{n_{(\ell)}+\tau_{(\ell)}}{n_{(j)}+\tau_{(j)}} = \left(\frac{\hat{\sigma}^2_{(\ell)}\zeta_{(\ell)}/c_{(\ell)}}{\hat{\sigma}^2_{(j)}\zeta_{(j)}/c_{(j)}}\right)^{1/2}. \tag{14}$$

The ratio in (14) is equivalent to the allocation in Step 4c of $\mathcal{LL}$ (cf. (8)) because $\zeta_{(j)} = \gamma_{(j)}$ for all j. Recall that the $\mathcal{LL}$ allocation is derived assuming $\tau \rightarrow \infty$ so that all systems with a nonzero variance will get some replications added (are in $\mathcal{S}$). The original derivation of Step 4c of $\mathcal{LL}$ in Chick and Inoue (2001) was for a two-stage procedure, but some algebra like that following (12) shows that the allocation holds for the sequential $\mathcal{LL}$ too. The basic insight is that both $\mathcal{LL}$ and $\mathcal{OCBA}_{\text{LL}}$ use Bonferroni-like bounds for the VIP and OCBA conceptualizations of EOC, respectively. Although $\mathcal{LL}$ uses $\lambda_{\{jk\}}$ and $\mathcal{OCBA}_{\text{LL}}$ uses $\tilde{\lambda}_{jk}$ in those bounds, they share a common limit, $\lim_{\tau\rightarrow\infty} \lambda_{\{jk\}} = \lambda_{jk} = \lim_{\tau\rightarrow 0} \tilde{\lambda}_{jk}$ in the sense described above. As a result the two procedures perform similarly (there are differences due to rounding when allocating).

We remark that (14) generalizes $\mathcal{OCBA}_{\text{LL}}$ to allow for different sampling costs for each system. It also permits $\mathcal{OCBA}_{\text{LL}}$ to be run rather flexibly as a two-stage procedure, just as for one of the $\mathcal{LL}$ variations in Chick and Inoue (2001): allocate some number B additional replications so that (14) holds, subject to a sampling budget constraint $\sum_{i=1}^{k} c_i\tau_i = B$.

C Computational Issues

The implementation that generated the analysis and graphs in this paper used the Gnu Scientific Library (gsl) for calculating cdfs, the Mersenne twister random number generator (Matsumoto and Nishimura 1998, with 2002 revised seeding), and FILIB++ (Lerch et al. 2001) for interval arithmetic. Calculations were run on a mixed cluster of up to 120 nodes. The nodes were running Linux 2.4 and Windows XP with Intel P4 and AMD Athlon processors ranging from 2 to 3 GHz. The program is written in C++ and jobs were distributed with the JOSCHKA-System (Bonn et al. 2005).

Numerical stability problems may arise in implementations of the OCBA allocations, even with double-precision floating point arithmetic, as the total number of replications gets quite large (i.e., for low values of α^* or EOC bounds). To better distinguish which system should receive samples in a given stage, numerical stability was increased by evaluating the system that maximizes $\log(\text{EAPCS}_i - \text{APCS})$ instead of EAPCS_i (e.g., for $\mathcal{OCBA}$) and $\log(\text{EEOCS}_i - \text{AEOC})$ (e.g., for $\mathcal{OCBA}_{\text{LL}}$). In particular, for $\mathcal{OCBA}$, set $p_j = \Phi_{\nu_{(j)(k)}}(d^*_{jk})$, $\tilde{p}_j = \Phi_{\tilde{\nu}_{(i)(k)}}(\tilde{\lambda}_{ik}^{1/2} d_{(i)(k)})$, for $\mathcal{OCBA}_{\text{LL}}$, set $q_i = \lambda_{ik}^{1/2}\Psi_{\nu_{(i)(k)}}(d^*_{ik})$, $\tilde{q}_i = \tilde{\lambda}_{ik}^{1/2}\Psi_{\tilde{\nu}_{(i)(k)}}(\tilde{\lambda}_{ik}^{1/2} d_{(i)(k)})$, and set $\sum_j = \sum_{j:(j)\neq(k)}$. Then

$$\log(\text{EAPCS}_i - \text{APCS})$$

$$= \begin{cases} \sum_j \log(1-p_j) - \log(1-p_i) + \log p_i + \log\left[-\left(\exp(\log\tilde{p}_i - p_i) - 1\right)\right] & \text{if } i \neq (k) \\ \sum_j \log(1-\tilde{p}_j) & \\ \quad + \log\left[-\left(\exp\left(\sum_j \log(1-p_j) - \sum_j \log(1-\tilde{p}_j)\right) - 1\right)\right] & \text{if } i = (k), \end{cases} \tag{15}$$

and

$$\log(\text{EEOCS}_i - \text{AEOC}) = \begin{cases} \log(q_i - \tilde{q}_i) & \text{if } i \neq (k) \\ \log(\sum_j q_j - \tilde{q}_j) & \text{if } i = (k). \end{cases} \tag{16}$$

These transformations are useful, because $\log(1-x)$ = `log1p(-x)` and $\exp(x) - 1$ = `expm1(x)` from the C runtime library have increased accuracy for x near 0. We used these transformations for the calculation of $1 - \text{PCS}_{\text{Slep}} = -(\exp(\sum_j \log(1-p_j)) - 1)$, too. In rare cases, we computed $\text{EAPCS}_i < \text{APCS}$ or $\text{EEOCS}_i < \text{AEOC}$, which we handled by setting $\log(\text{EAPCS}_i - \text{APCS})$ to $-\infty$.

For calculating $\log p_j$ and $\log q_j$, we need $\log\Phi_\nu(t)$ and $\log\Psi_\nu(t)$. If the numerical stability does not suffice to calculate $\log\Phi_\nu(t)$ (underflow error) we derive bounds for $\log\Phi_\nu(t)$ based on the following property of the cdf of a t-distribution (Evans, Hastings, and Peacock 1993),

$$\Phi_\nu(t) = \begin{cases} \frac{1}{2}\beta^{\text{inc}}_{\text{reg}}(\frac{\nu}{2}, \frac{1}{2}, \frac{\nu}{\nu+t^2}) & \text{if } t \leq 0 \\ 1 - \frac{1}{2}\beta^{\text{inc}}_{\text{reg}}(\frac{\nu}{2}, \frac{1}{2}, \frac{\nu}{\nu+t^2}) & \text{if } t > 0, \end{cases} \tag{17}$$

where $\beta^{\text{inc}}_{\text{reg}}(a,b,x) = \beta(a,b)^{-1}\int_0^x u^{a-1}(1-u)^{b-1}du$ is the incomplete beta function, and $\beta(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ denotes the beta function. A lower bound for $\log\Phi_\nu(-t)$ for $t > 0$ can be derived as follows. If $f(u) = u^{a-1}(1-u)^{b-1}$, then $f(0) = 0$, $f'(u) \geq 0$ and $f''(u) > 0$ for $a = \frac{\nu}{2} > 1$, $b = \frac{1}{2}$ and all $u \in [0,1]$. So the area below $f(u)$ over $[0,x]$ is always larger than the area below the tangent at $(x, f(x))$.

$$\log\Phi_\nu(-t) \geq \frac{\nu}{2}\log\frac{\nu}{t^2+\nu} + \frac{1}{2}\log(1 - \frac{\nu}{t^2+\nu}) - \log\left((\frac{\nu}{2} - 1)(1 - \frac{\nu}{t^2+\nu}) + \frac{1}{2}\frac{\nu}{t^2+\nu}\right) - \log 2 \tag{18}$$

For the upper bound, recall (6). As $\Psi_\nu(t) > 0$ for all $t > 0$ we obtain $\Phi_\nu(-t) < \frac{1}{t}\frac{\nu+t^2}{\nu-1}\phi_\nu(t)$, so

$$\log \Phi_\nu(-t) \quad < \quad \log \tfrac{\nu/t+t}{\nu-1} + \log \phi_\nu(t). \tag{19}$$

Note that $\log \phi_\nu(t)$ can be calculated by using the logarithm of the Gamma-function,

$$\log \phi_\nu(t) \quad = \quad \log \Gamma\left(\tfrac{\nu+1}{2}\right) - \log \Gamma\left(\tfrac{\nu}{2}\right) - \tfrac{1}{2}\log(\nu\pi) - \tfrac{\nu+1}{2}\log\left(1 + \tfrac{t^2}{\nu}\right). \tag{20}$$

Collisions, due to $\log(\mathrm{EAPCS}_i - \mathrm{APCS})$ or $\log(\mathrm{EEOCS}_i - \mathrm{AEOC})$ being not numerically unique because of the interval bounds above, occurred rarely with $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\mathrm{LL}}$. If there was no clearly defined best and $\mathrm{EAPCS}_i - \mathrm{APCS}$ or $\mathrm{EEOCS}_i - \mathrm{AEOC}$ was not numerically different from 0 for *any* system (with interval arithmetic), then we repeatedly doubled τ for purposes of calculating $\mathrm{EAPCS}_i - \mathrm{APCS}$ and $\mathrm{EEOCS}_i - \mathrm{AEOC}$, until at least one system was numerically greater than 0. The 'winner' then received 1 replication. Usually, doubling at most 3 times ($\tau = 8$) was sufficient to select a winner. If there was no clearly defined best because two or more systems whose $\mathrm{EAPCS}_i - \mathrm{APCS}$ or $\mathrm{EEOCS}_i - \mathrm{AEOC}$ had overlapping intervals but the intervals did not contain 0, then we allocated $\tau = 1$ replication to the system with the highest upper bound for the interval. Because the interval arithmetic increased CPU time by 50 % and we did not observe different allocations for $\mathcal{OCBA}$ when resolving collisions as described or by simply using the upper bounds, we ran the experiments with the upper bounds.

Although collisions occur rarely with $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\mathrm{LL}}$, there is some slight bend to the right for low values of α^* or EOC bounds. That may suggest a potential inefficiency due to another numerical issue that we have not yet identified. Increasing τ_0 can help to reduce the bending for low values of α^* or EOC.

Numerical stability problems also arise with some other allocations that we tested. Those other allocations are not presented in this paper, as they were less effective than the allocations in the main paper, in spite of the assistance that they received from the above ideas that are designed to improve numerical stability.

In conclusion, we tested a number of numerical techniques in order to improve the performance of each procedure. Some, but not all, of those numerical techniques required additional computational overhead in order to compute the allocation. Monte Carlo estimates and complicated quadrature, the techniques that required the most additional overhead, did not necessarily help more. The procedures that improved the most from the help from interval arithmetic were not the best procedures and are not reported here any more due to size restrictions. The $\mathcal{OCBA}$ and $\mathcal{OCBA}_{\mathrm{LL}}$ required some attention for numerical stability, but not much, and without any notable decrease in the time required to compute an allocation.

Procedures $\mathcal{KN}$++, $\mathcal{LL}$, and 0-1 did not experience numerical stability problems with collisions.

D Design Settings for Experimental Test Bed

This experiment appears to have tested more settings for the sampling allocations, stopping rules, number of systems, and selection problem configurations than any previous empirical study.

The study served to identify consistent patterns as a function of the number of systems, the difficulty of the problem structure (as identified by either the closeness of the means of the competitors for the best; or by the size of the variances of the different alternatives), the number of first stage replications, the stopping rules, the allocations, and so forth. When we observed clear trends in multiple scenarios (e.g., sensitivity to the number of first stage replications, etc.), we documented the result. When there was an unclear pattern, we ran many additional replications in order to clarify the mixed message.

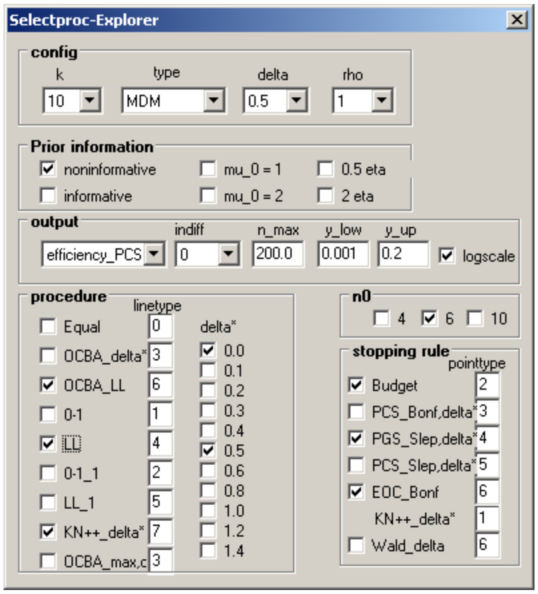
The experimental design therefore represents the testing of multiple first-order trends for varying problem structures, with additional exploration in ‘interesting’ areas.

The quantification of interactions between two parameters, say the number of first-stage replications and the number of systems, was not a primary goal. Such a quantification with a linear response model would likely be subject to biases anyway. Since the use of any given parametric response model would likely give model mis-specification errors, we did not propose one. We therefore did not use experimental design techniques (such as fractional factorial) to select settings. We note that experimental design techniques are not typically used in such settings.

Our choice of structured configurations is similar in spirit to the approach of Nelson, Swann, Goldsman, and Song (2001), which is one of the larger empirical studies that we have seen published. That study compared several IZ procedures.

Table 3 gives a listing of the configurations that were tested. Not all procedures were tested in all settings. When $k \geq 50$, it often turned out that some procedures would not finish running in a reasonable time (e.g., overly conservative procedures for those settings). So we mostly explored configurations with $k \leq 20$. We ran $\mathcal{S}$ in many but not all settings, since its behavior became clear part way through the study. Similarly, we tested the effects of n_0 on many configurations, but once we determined that $n_0 < 6$ behaved poorly, we focused on $n_0 = 6$ for most cases, and tested $n_0 = 10$ on a subset of configurations for additional sensitivity. We only tested the inclusion of prior information for the VIP and OCBA on a handful of settings.

In spite of this pruning, we arrived at 25,000+ different combinations of configurations and parameters for the procedures (allocations, stopping rules, n_0, δ^*, α^*, β^*, etc., including some allocations and stopping rules that were reported in preliminary work of Branke, Chick, and Schmidt (2005) or small-sample derivations for VIP procedures, and that were tested on a subset of configurations, but were dropped due to lower performance). That must be multiplied again to account for the multiple types of graphs that we considered (PBS$_{\mathrm{IZ},\delta^*}$ and EOC$_{\mathrm{IZ}}$, both efficiency and target). With so many combinations, we cannot claim that we examined each and every one visually with our output browser. However, we can say that we examined at least 10,000 of them. One can view the output of multiple procedures simultaneously on a single plot (up to 10 or 15, reasonably), can browse from one parameter setting to another very quickly using the interface in Figure 45, and three coauthors can work in parallel. We systematically varied each main parameter that was discussed in the paper over a variety of configurations and parameter settings to insure that any general claims (not specific to particular setting) in the main paper represent qualitative conclusions that we viewed repeatedly for a number of settings.



Selectproc-Explorer

config

k: 10 ▾ type: MDM ▾ delta: 0.5 ▾ rho: 1 ▾

Prior information

☑ noninformative ☐ mu_0 = 1 ☐ 0.5 eta

☐ informative ☐ mu_0 = 2 ☐ 2 eta

output

efficiency_PCS ▾ indiff: 0 ▾ n_max: 200.0 y_low: 0.001 y_up: 0.2 ☑ logscale

procedure

	linetype
☐ Equal	0
☐ OCBA_delta*	3
☑ OCBA_LL	6
☐ 0-1	1
☑ LL	4
☐ 0-1_1	2
☐ LL_1	5
☑ KN++_delta*	7
☐ OCBA_max,c	3

delta*

☑ 0.0
☐ 0.1
☐ 0.2
☐ 0.3
☐ 0.4
☑ 0.5
☐ 0.6
☐ 0.8
☐ 1.0
☐ 1.2
☐ 1.4

n0

☐ 4 ☑ 6 ☐ 10

stopping rule

	pointtype
☑ Budget	2
☐ PCS_Bonf,delta*	3
☑ PGS_Slep,delta*	4
☐ PCS_Slep,delta*	5
☑ EOC_Bonf	6
KN++_delta*	1
☐ Wald_delta	6

Figure 45: Graphical User Interface for Browsing and Visualizing Selection Procedure Performance.

REFERENCES

Bonn, M., F. Toussaint, and H. Schmeck (2005). JOSCHKA: Job-Scheduling in heterogenen Systemen mit Hilfe von Webservices. In E. Maehle (Ed.), *PARS Workshop Lübeck*, pp. 99–106. Gesellschaft für Informatik.

Chick, S. E. and K. Inoue (2001). New two-stage and sequential procedures for selecting the best simulated system. *Operations Research 49*(5), 732–743.

Evans, M., N. Hastings, and B. Peacock (1993). *Statistical Distributions* (2 ed.). New York: Wiley.

Lerch, M., G. Tischler, J. W. von Gudenberg, W. Hofschuster, and W. Kraemer (2001). The interval library filib++ 2.0 - design, features and sample programs. Preprint 2001/4, University of Wuppertal.

Matsumoto, M. and T. Nishimura (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM TOMACS 8*(1), 3–30.

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.

```
  Slippage Configuration: <# systems>_SC_<delta>_<rho>
10_SC_0.25_1/     20_SC_0.354_1/    2_SC_0.25_1/     50_SC_0.354_1/
10_SC_0.354_1/    20_SC_0.5_0.354/  2_SC_0.354_1/    50_SC_0.5_1/
10_SC_0.5_0.125/  20_SC_0.5_0.5/    2_SC_0.5_0.125/  50_SC_0.707_1/
10_SC_0.5_0.177/  20_SC_0.5_0.707/  2_SC_0.5_0.177/  50_SC_1_1/
10_SC_0.5_0.25/   20_SC_0.5_1.414/  2_SC_0.5_0.25/   5_SC_0.25_1/
10_SC_0.5_0.354/  20_SC_0.5_1.5/    2_SC_0.5_0.354/  5_SC_0.354_1/
10_SC_0.5_0.5/    20_SC_0.5_1/      2_SC_0.5_0.5/    5_SC_0.5_0.354/
10_SC_0.5_0.707/  20_SC_0.5_2.828/  2_SC_0.5_0.707/  5_SC_0.5_0.5/
10_SC_0.5_1.414/  20_SC_0.5_4/      2_SC_0.5_1.414/  5_SC_0.5_0.707/
10_SC_0.5_1.5/    20_SC_0.707_1/    2_SC_0.5_1.5/    5_SC_0.5_1.414/
10_SC_0.5_1/      20_SC_1_1/        2_SC_0.5_1/      5_SC_0.5_1.5/
10_SC_0.5_2.828/  2_SC_0.0625_1/    2_SC_0.5_2.828/  5_SC_0.5_1/
10_SC_0.5_4/      2_SC_0.125_1/     2_SC_0.5_4/      5_SC_0.5_2.828/
10_SC_0.707_1/    2_SC_0.177_1/     2_SC_0.707_1/    5_SC_0.5_4/
10_SC_1_1/        2_SC_0.25_0.354/  2_SC_1_1/        5_SC_0.707_1/
20_SC_0.25_1/     2_SC_0.25_0.5/    50_SC_0.25_1/    5_SC_1_1/
  Monotone Decreasing Means Configurations: <# systems>_MDM_<delta>_<rho>
100_MDM_0.1_1/     10_MDM_0.5_2.828/  2_MDM_0.354_1/      50_MDM_0.707_1/
10_MDM_0.118_1/    10_MDM_0.5_4/      2_MDM_0.5_0.00195/  50_MDM_1_1/
10_MDM_0.165_1/    10_MDM_0.707_0.5/  2_MDM_0.5_0.00781/  5_MDM_0.25_0.354/
10_MDM_0.167_1/    10_MDM_0.707_1/    2_MDM_0.5_0.354/    5_MDM_0.25_0.5/
10_MDM_0.25_0.5/   10_MDM_1.414_0.5/  2_MDM_0.5_0.5/      5_MDM_0.25_1/
10_MDM_0.25_1/     10_MDM_1_0.5/      2_MDM_0.5_0.707/    5_MDM_0.354_1/
10_MDM_0.354_0.5/  10_MDM_1_1/        2_MDM_0.5_1.414/    5_MDM_0.5_0.25/
10_MDM_0.354_1/    10_MDM_2_0.5/      2_MDM_0.5_1.5/      5_MDM_0.5_0.354/
10_MDM_0.5_0.125/  20_MDM_0.25_1/     2_MDM_0.5_1/        5_MDM_0.5_0.5/
10_MDM_0.5_0.177/  20_MDM_0.354_1/    2_MDM_0.5_2.828/    5_MDM_0.5_0.707/
10_MDM_0.5_0.25/   20_MDM_0.5_1/      2_MDM_0.5_4/        5_MDM_0.5_1.414/
10_MDM_0.5_0.354/  20_MDM_0.707_1/    2_MDM_0.707_1/      5_MDM_0.5_1.5/
10_MDM_0.5_0.5/    20_MDM_1_1/        2_MDM_1_1/          5_MDM_0.5_1/
10_MDM_0.5_0.707/  2_MDM_0.04_1/      3_MDM_0.5_0.125/    5_MDM_0.5_2.828/
10_MDM_0.5_1.414/  2_MDM_0.25_0.354/  50_MDM_0.25_1/      5_MDM_0.5_4/
10_MDM_0.5_1.5/    2_MDM_0.25_0.5/    50_MDM_0.354_1/     5_MDM_0.707_1/
10_MDM_0.5_1/      2_MDM_0.25_1/      50_MDM_0.5_1/       5_MDM_1_1/
  Random Problem Instances, RPI1: <# systems>_RPI_<eta>_<b>
10_RPI_0.5_100/    20_RPI_2_100/     5_RPI_0.354_100/   5_RPI_2.828_100/
10_RPI_0.707_100/  2_RPI_0.707_100/  5_RPI_0.5_100/     5_RPI_2_100/
10_RPI_1.414_100/  2_RPI_1.414_100/  5_RPI_0.707_100/   5_RPI_2_2.5/
10_RPI_1_100/      2_RPI_1_100/      5_RPI_1.414_100/   5_RPI_4_100/
10_RPI_1_2.5/      2_RPI_1_2.5/      5_RPI_11.314_100/  5_RPI_5.657_100/
10_RPI_2_100/      2_RPI_2_100/      5_RPI_16_100/      5_RPI_8_100/
10_RPI_2_2.5/      2_RPI_2_2.5/      5_RPI_1_100/
20_RPI_1_100/      50_RPI_2_100/     5_RPI_1_2.5/
  Random Problem Instances, RPI2: <# systems>_RPI3_<eta>_<b>
  (an extra - before <eta> in the name indicates a=0; otherwise a=1)
10_RPI3_1.414_100/  2_RPI3_-1_100/     2_RPI3_2_2.5/      5_RPI3_1_100/
10_RPI3_1.414_2.5/  2_RPI3_1.414_100/  5_RPI3_-1_100/     5_RPI3_1_2.5/
10_RPI3_1_100/      2_RPI3_1.414_2.5/  5_RPI3_0.5_100/    5_RPI3_2_100/
10_RPI3_1_2.5/      2_RPI3_1_100/      5_RPI3_0.707_100/  5_RPI3_2_2.5/
10_RPI3_2_100/      2_RPI3_1_2.5/      5_RPI3_1.414_100/
10_RPI3_2_2.5/      2_RPI3_2_100/      5_RPI3_1.414_2.5/
```

Table 3: Configurations Tested in the Empirical Study. §3 Defines the Configuration's Parameters.