J. Statist. Comput. Simul., 1986, Vol. 23, pp. 183–209
0094-9655/86/2303-0183 $18.50/0

Printed in Great Britain

Decomposition of Some Well-Known Variance Reduction Techniques†

BARRY L. NELSON

Department of Industrial and Systems Engineering, The Ohio Sta University, Columbus, OH 43210, U.S.A.

and

BRUCE W. SCHMEISER

School of Industrial Engineering, Purdue University, West Lafayette, I 47907, U.S.A.

(*Received December 16, 1984; in final form August 19, 1985*)

Proliferation of techniques and lack of unifying structure have hindred both study and application of variance reduction. We view any variance reduction technique (VRT) as a transformation from one experiment to another, which leads to an exhaustive taxonomy of VRTs. In this paper, seven familiar VRTs are expressed in our taxonomy as compositions of elemental transformations from six basic classes. Our objective is to use the reader's knowledge of these well-known techniques to illustrate the taxonomy.

KEY WORDS: Antithetic variates, common random numbers, conditional expectations, control variates, importance sampling, Monte Carlo, poststratified sampling, simulation, swindles, variance reduction.

---

†Research partially supported by Office of Naval Research contract N00014-79-C-0832.

1. INTRODUCTION

A computer simulation model of a real or conceptual system consists of a probabilistic representation of those elements in the system that cannot be predicted with certainty and deterministic rules that define the system's reaction to realizations of the uncertain elements. For instance, in a model of a queueing system the time between arrivals is uncertain while the queue discipline specifies a rule for processing customers after arrival. Such a model mimics the actual functioning of the system, or at least the essential elements of it. A simulation experiment is performed by generating realizations of the model and computing estimates of performance measures.

Variance Reduction Techniques (VRTs) are transformations. They transform a simulation experiment into a related experiment that yields better estimators of the performance measures, where better means more precise. This gain in precision may be at the expense of the one-to-one correspondence between the model and the system. See Bratley, Fox and Schrage (1983) and Wilson (1983a, 1985) for surveys of variance reduction.

In Nelson and Schmeiser (1984b) we identify six classes of transformations that exhaust the set of all possible variance reduction techniques under composition. The derivation of the six classes is based on a mathematical-statistical definition of simulation experiments developed specifically for studying variance reduction (Nelson and Schmeiser, 1984a). We are proposing a taxonomy for variance reduction.

Given any taxonomy of variance reduction, what should it do to be useful? Certainly it should unify the field as it currently exists and also bring new insights into what could potentially exist. To be specific, we think a taxonomy should: (a) eliminate confusion regarding the characteristics of, and relationships among, VRTs; (b) provide a common language for communication in practice, research, and instruction; (c) help practitioners determine appropriate variance reduction strategies; (d) generate new variance reduction ideas; and (e) provide a basis for automation of variance reduction.

We return to these five criteria in Section 5 after illustrating the taxonomy in Section 4. The examples of Section 4 are based on graphical symbols presented in Section 3 and definitions of simulation experiment and of the six classes of transformations presented in Section 2.

2. BACKGROUND

Descriptions are in terms of matrices, columns of matrices, and elements of matrices. Letters, Greek or Roman, without subscripts denote matrices, letters with single subscripts denote columns, and doubly subscripted letters are scalar elements, using the usual row-column convention. For instance, X_{ik} is the ith element of column vector X_k, which is the kth column in the matrix X. For our purposes, a matrix need not have elements in all positions, since elements of sets are arranged in rows and columns for conceptual rather than computational reasons.

A letter with subscripts in parentheses indicates a set of variables with indices in a fixed set. For example, $X_{(ab)}$ denotes all elements X_{ik} in X with subscripts in index set (ab), a set that would have to be defined. Thus $(.)$ is a mapping from a single index to a set of indices.

Random variables are denoted by capital Roman letters, and realizations of these random variables by lower case letters. For example, y_{il} is a realization of random variable Y_{il}. Any notation that is counter to the above conventions is specifically defined as needed.

For our purposes, a *simulation experiment* is a description of a system of random variables; the dimensions of this system may also be random variables. Given a source of randomness (usually independent $U(0, 1)$ random variables), realizations of the system can be generated. Based on the definitions in Nelson and Schmeiser (1984a), the random variables are partitioned into *inputs*, *outputs* and *statistics*. These sets have precise definitions, but can be described loosely as follows:

Inputs are random variables defined by known (although possibly only conditionally) probability distributions. Examples are service and interarrival times of a queueing simulation, or the demand per period of an inventory system. Another example is a service time whose distribution, conditional on the number of customers in the system, is known. The countably infinite matrix of inputs is denoted by X and has joint cumulative distribution function $F(x)$.

Outputs are random variables defined by known, deterministic functions of the inputs. They are the observations of system performance, such as the delay experienced by customers in the queueing simulation or the number of backorders in the inventory system. The

probability distribution of the outputs is not known, but the functions define how outputs are realized from the inputs. The output, Y, is the matrix of all *essential* random variables defined by functions of X, in the sense that all remaining random variables that are functions of X can be derived from Y, provided no element of Y is deleted. The outputs are denoted by $Y=g(X;R_*)$, where R_* is the sampling plan; the sampling plan defines a stopping rule for the simulation experiment in terms of the number of realizations in columns of Y.

Statistics are functions that aggregate outputs into point estimators of the performance measures of interest. A sample mean is often used. Variance reduction refers to reducing the variance of these statistics. The statistics are denoted by $Z=h(Y)$, and the performance measures of interest by θ; Z and θ are row vectors of the same dimension.

If Z and θ are scalars, and Z is an unbiased estimator of θ, then

$$\mathrm{Var}(Z)=\int[h(g(x;R_*))-\theta]^2\,dF(x)$$

VRTs are transformation of simulation experiments that alter the inputs, outputs and statistics to reduce $\mathrm{Var}(Z)$. If we hold θ and the sample space of X fixed, then variance reduction must be accomplished by redefining F, g, R_*, and/or h. Six classes of transformations, defined loosely here, that exhaust the possibilities are:

—*Distribution Replacement* (DR): Redefine the scalar marginal distributions of the inputs without altering any statistical dependencies among the inputs.

—*Dependence Induction* (DI): Redefine the statistical dependencies among the scalar inputs without altering any marginal distributions of the inputs.

—*Equivalent Allocation* (EA): Redefine the functions from inputs to outputs without altering the allocation of sampling effort.

—*Sample Allocation* (SA): Redefine the allocation of sampling effort without altering the functions that define the outputs.

—*Equivalent Information* (EI): Redefine the functions from outputs to statistics without altering the argument set of the statistics.

—*Auxiliary Information* (AI): Redefine the argument set of the statistics without altering the functions from outputs to statistics.

Transformations in DR and DI redefine F; those in EA and SA redefine g and R_*, respectively; and those in EI and AI redefine h. The mathematical rigor needed to prove properties such as exhaustiveness is presented in Nelson and Schmeiser (1984a, b); precise definitions are needed to make the partitioning of inputs, outputs and statistics unambiguous, and to make the classes of transformations distinct. However, the loose, intuitive definitions given here are more useful for our current purpose.

3. SYMBOL SET

To augment the discussion of the seven VRTs in Section 4, a graphical representation of each VRT is given. Only three symbols are needed, as illustrated in Figure 1. Rectangles enclose a definition of an input, output, or statistic in the simulation experiment. Circles enclose a class of transformations. Trapezoids contain the prior knowledge used to make application of the transformation possible and reasonable.

By *prior knowledge* we mean any knowledge, either known with certainty or suspected, beyond that necessary to design the original

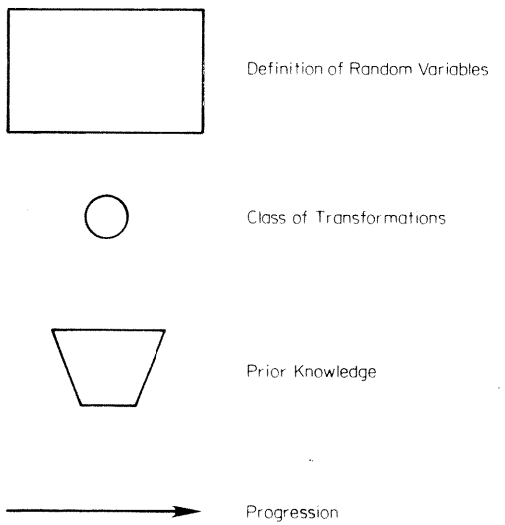

FIGURE 1 Symbol set for VRT decomposition.

("crude") simulation experiment. Since prior knowledge is often difficult to state succinctly, and since our purpose here is to illustrate the taxonomy rather than to provide every detail of the VRTs discussed, the prior knowledge specified in Figures 2–8 is often minimal. Complete specification is often clear by considering implementation of the VRT.

The progression in each figure is from left to right, proceeding from a definition of some input, output, or statistic to a new definition via a transformation.

4. DECOMPOSITION OF VARIANCE REDUCTION TECHNIQUES

The VRTs considered are antithetic variates, common random numbers, control variates, stratified sampling, poststratifying the sample, conditional expectations, and importance sampling. These seven were selected specifically because they are well-known and understood, and thus they provide a convenient introduction to our variance reduction taxonomy. For each VRT, a description of the VRT and graphical display of its decomposition is presented. The purpose is not to propose a comprehensive definition of these seven VRTs, but rather to illustrate our taxonomy using well-known examples.

Each transformation in the decomposition of a VRT may not, by itself, reduce variance or even yield an acceptable experiment. Variance reduction is achieved by the combined effect of all the relevant transformations.

4.1 Antithetic Variates (AV)

Antithetic Variates is a VRT that has been extensively studied in the context of Monte Carlo estimation. The technique first appeared in Hammersley and Morton (1956), with further early developments by Hammersley and Mauldon (1956), Morton (1957), Halton and Handscomb (1957), and Handscomb (1958). In its broadest sense, "we use the term antithetic variates to describe any set of estimators which mutually compensate each other's variations" (Hammersley

and Handscomb, 1964, p. 61). Statistical results such as

$$\mathrm{Var}(Y_i \pm Y_j) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_j) \pm 2\ \mathrm{Cov}(Y_i, Y_j) \tag{1}$$

indicate the advantage of forcing correlation among outputs while maintaining their marginal distributions. Antithetic variates attempts to induce negative correlations among identically distributed simulation outputs.

Consider estimating θ_1, a real scalar, using a simulation experiment defining

$$Y_{i1} = g_{i1}(X_{(i1)}) \quad i = 1, 2, \ldots, I_1 = 2n$$

where $E(Y_{i1}) = \theta_1$ and $X_{(i1)}$ is a set of inputs indexed by i, with statistic

$$Z_1 = I_1^{-1} \sum_{i=1}^{I_1} Y_{i1}$$

Further, suppose that

$$X_{(i1)} \sim \text{i.i.d. } F_{(1)}(x_{(i1)}) \quad i = 1, 2, \ldots, 2n \tag{2}$$

The usual AV transformation is to redefine the joint distribution of

$$(X_{(2i-1,1)}, X_{(2i,1)}) \quad i = 1, 2, \ldots, n$$

such that the marginals given in (2) are unchanged, but the pairs are negatively correlated. When $X_{(i1)}$ is a scalar, or if AV is used only on a scalar component of $X_{(i1)}$, the correlation is most often induced by generating realizations via the inverse cumulative distribution function (cdf) of $X_{(i1)}$ in the following manner:

$$X_{(2i-1,1)} = F_{(1)}^{-1}(U_i)$$

$$X_{(2i,1)} = F_{(1)}^{-1}(1 - U_i)$$

where $U_i \sim \text{i.i.d. } U(0,1)$ $i = 1, 2, \ldots, n$; this induces the minimal achievable covariance between the inputs $X_{(2i-1,1)}$ and $X_{(2i,1)}$ with

the given marginal distributions (Whitt, 1976). The resulting joint cdf is

$$\max\{F_{(1)}(x_{(2i-1,1)})+F_{(1)}(x_{(2i,1)})-1,\ 0\}$$

Thus, AV is composed of a single transformation from the Dependence Induction (DI) class, as shown in Figure 2.

When correlation is to be induced among k-tuples ($k>2$), there are a variety of approaches and objectives (see for instance, Fishman and Huang, 1983). Although more complicated mathematically, the k-tuple case still involves only a single transformation from DI.

The reason for inducing dependence among the inputs is to cause

$$\mathrm{Cov}(Y_{2i-1,1}, Y_{2i,1})<0$$

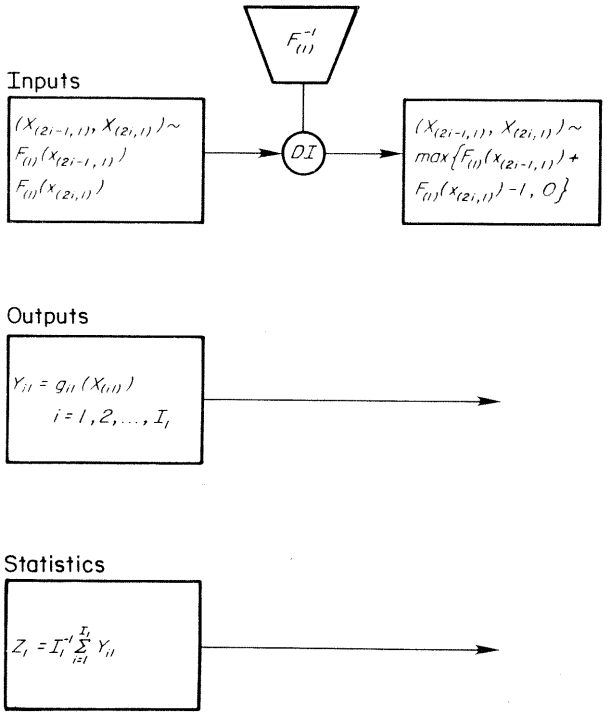

FIGURE 2 Antithetic variates (AV).

which reduces the variance of Z_1 via (1). Negative covariance between the outputs is not guaranteed by negative covariance between the inputs. However, Wilson (1983) showed that if the inputs are generated via the inverse cdf (as shown above) and g_{i1} is monotone, then the minimum achievable covariance between $Y_{2i-1,1}$ and $Y_{2i,1}$ will be achieved. However, achieving the bound, or even reducing variance at all, is not a factor in classifying these methods as AV or the transformation as DI. The only relevant factor is that the correlation is induced by redefining the inputs while maintaining their marginal distributions.

This redefinition is usually accomplished by making the inputs deterministically dependent. When an estimator consists of a sum of n outputs, the *antithetic-variates theorem* (Hammersley and Mauldon, 1956, Handscomb, 1958, Wilson, 1979, 1983) states that under fairly general conditions the greatest lower bound of the variance of the estimator can be approached arbitrarily closely by generating all n inputs from a deterministic transformation of a single randomly sampled input. But whether or not correlation is induced deterministically, the transformation is still DI in our taxonomy.

4.2 Common Random Numbers (CRN)

Common random numbers is often called *correlated sampling* (CS). Confusion can arise because CRN is both a method for generating correlated samples and a VRT that exploits induced correlation. "The name of the technique stems from the possibility in some situations of using the *same* stream of basic $U(0,1)$ random variables to drive each of the alternative models through time..." (Law and Kelton, 1982, p. 350). We use the term CRN in the sense of CS, meaning that correlation is induced (by whatever means) between certain inputs to obtain positively correlated outputs for estimating the difference between two performance measures. CRN has the distinction of being "...the only VRT that is as a rule used by practitioners of simulation" (Kleijnen, 1974, p. 206).

Consider estimating

$$\theta_1=\alpha_2-\alpha_3$$

where α_2 and α_3 are real, scalar constants, using a simulation

experiment defining

$$Y_{il}=g_{il}(X_{(il)}) \quad i=1, 2,\ldots, I_l \quad l=2, 3$$

where $E(Y_{il})=\alpha_l$, with statistic

$$Z_1=I_2^{-1}\sum_{i=1}^{I_2} Y_{i2}-I_3^{-1}\sum_{i=1}^{I_3} Y_{i3}=\bar{Y}_2-\bar{Y}_3$$

The basis for CRN is the well-known relation

$$\mathrm{Var}(\bar{Y}_2-\bar{Y}_3)=\mathrm{Var}(\bar{Y}_2)+\mathrm{Var}(\bar{Y}_3)-2\,\mathrm{Cov}(\bar{Y}_2,\bar{Y}_3)$$

Aggregating the individual inputs X_{il} into two sets of inputs corresponding to the two systems, we can write

$$Y_l=g_l(X_{(l)}) \quad l=2, 3$$

which defines two aggregated sets of outputs. The original experiment typically has $X_{(2)}$ and $X_{(3)}$ independent; that is, the two systems are realized using different sequences of $U(0,1)$ random numbers. CRN redefines the joint distribution of $(X_{(2)}, X_{(3)})$, without changing their multivariate marginal distributions, in a way the practitioner hopes will induce $\mathrm{Cov}(\bar{Y}_2,\bar{Y}_3)>0$ and in turn a variance reduction for Z_1. Thus, CRN consists of a single transformation from the DI class.

If pairs of scalar inputs, say (X_{i2}, X_{i3}), can be identified such that each pair is independent of all other pairs, $X_{i2}\in X_{(2)}$, and $X_{i3}\in X_{(3)}$, then positive correlation can be induced within each pair by generating observations with the same $U(0,1)$ random number sequence using the inverse cdf

$$X_{il}=F_{il}^{-1}(U_i) \quad l=2,3$$

which results in the maximum achievable $\mathrm{Cov}(X_{i2}, X_{i3})$ and joint cdf

$$\min\{F_{i2}(x_{i2}), F_{i3}(x_{i3})\}$$

This simple version of CRN is shown in Figure 3.

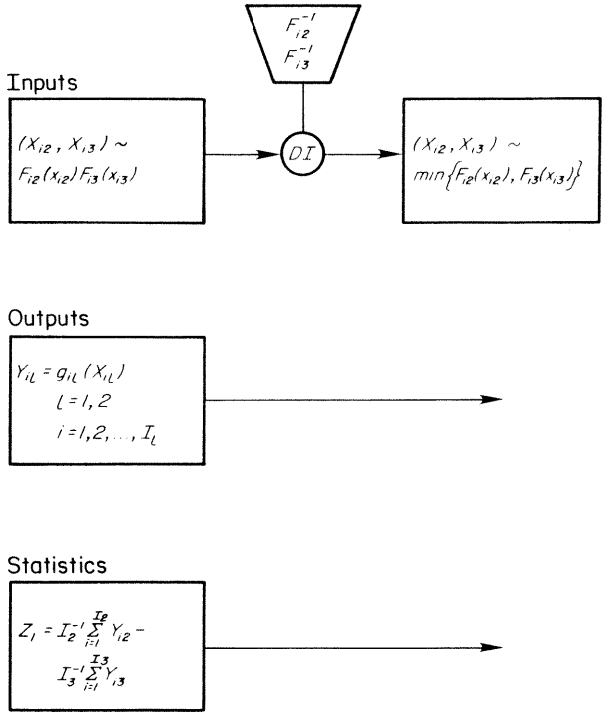

FIGURE 3 Common random numbers (CRN).

The desired positive correlation between the outputs is not guaranteed merely by inducing positive correlation between the inputs. However, analogous to antithetic variates, if the inverse cdf is used to generate the inputs, then monotonicity of the g_{il} functions ensures a favorable covariance term. Here again, whether a variance reduction is achieved is not relevant to the decomposition of CRN. Similarly, the decomposition is the same when the inputs are an historical trace or when nondeterministic methods (such as blocking in the experimental design) are used.

4.3 Control Variates (CV)

By the term *control variates* we refer to statistics that attempt to correct the value of an estimator based on the discrepancy between

the value of a second estimator and the known value of its expectation. For example, let $Y_{(1)}$ and $Y_{(2)}$ be sets of output random variables in a simulation experiment, and s_1 and s_2 be known scalar-valued functions such that

$$E[s_1(Y_{(1)})]=\theta_1 \quad \text{and} \quad E[s_2(Y_{(2)})]=\alpha$$

where θ_1 and α are real scalars; θ_1 is the performance measure of interest and α is known. The two most common CV estimators of θ_1 are the linear control

$$Z_c=s_1(Y_{(1)})-b(s_2(Y_{(2)})-\alpha) \tag{3}$$

where b is a scalar, and the ratio estimator

$$Z_c=\frac{s_1(Y_{(1)})}{s_2(Y_{(2)})}\alpha \tag{4}$$

The function s_2 is the control variate.

Both (3) and (4) are of the form

$$Z_c=h(s_1(Y_{(1)}),s_2(Y_{(2)})) \tag{5}$$

with the property that $h(\theta_1,\alpha)=\theta_1$. Several authors have noted that these two estimators are similar, including Kleijnen (1974) and Isaki (1983).

As shown in Figure 4, statistics such as (5) are obtained by a composite transformation that first augments the argument with output $Y_{(2)}$, which is AI, then modifies the statistic h, which is EI.

Both (3) and (4) extend naturally to multiple control variates, which does not change the decomposition. Whether b is a constant or is estimated from the outputs also does not change the decomposition.

In the simulation literature, a distinction is made between "internal" control variates (random variables that are part of the same real or conceptual system) and "external" control variates (random variables that are part of a similar real or conceptual system). This distinction is important in our taxonomy. Internal CV, shown in Figure 4, makes use of inherent correlation within the single system.

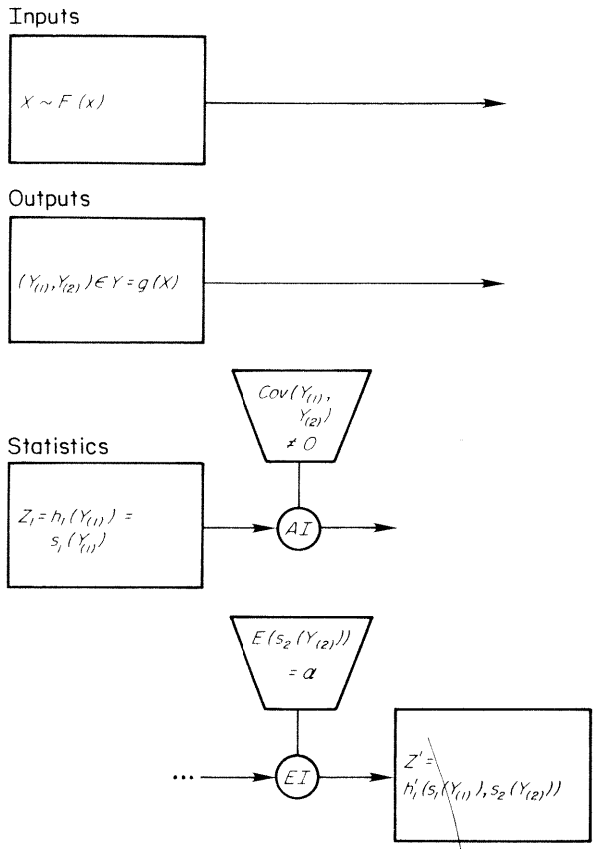

FIGURE 4 Internal control variates (CV).

However, external control variates employ an additional DI transformation to induce statistical dependence between the two systems; as in the CRN example (Section 4.2), both systems are part of the same simulation experiment in our definition.

4.4 Stratified Sampling (STRAT)

STRAT is a technique that replaces simple random sampling with a sampling plan designed to reduce variance. Hammersley and Handscomb (1964) and Rubinstein (1981) discuss stratified sampling

in the context of Monte Carlo problems and Cochran (1977) discusses the context of survey sampling. Books containing chapters dealing with stratified sampling specifically in systems simulation are Kleijnen (1974) and Bratley, Fox and Schrage (1983).

Consider estimating θ_1 when it is possible to sample I_1 observations of Y_{i1}, where $E(Y_{i1})=\theta_1$, $i=1, 2,\ldots, I_1$. The crude estimator of θ_1 might be

$$Z_1=h_1(Y_{(1)})=I_1^{-1}\sum_{i=1}^{I_1} Y_{i1}$$

Now suppose Y_{i1} can be expressed as a function of $(X_{(i1)}, X_{ik})$ for some fixed column index k of X. For notational convenience write

$$Y_{i1}=g_{i1}(X_{ik}) \tag{6}$$

suppressing the $X_{(i1)}$. Assume that X_{ik} are i.i.d. random variables for all i, and that the range of X_{ik} can be divided into n nonoverlapping, exhaustive strata (intervals). Denote these strata by L_j, $j=2,3,\ldots,n+1$. An equivalent way to view (6) is

$$Y_{mj}=g_{mj}(X_k) \quad j=2,3,\ldots,n+1 \quad m=1,2,\ldots,I_j$$

such that Y_{mj} is the mth observation of Y_1 for which the associated random variable $X_{ik}\in L_j$, and

$$\sum_{j=2}^{n+1} I_j=I_1$$

The STRAT estimator is

$$Z_1'=h_1'(Y_{(1)})=\sum_{j=2}^{n+1} p_j\left(I_j'^{-1}\sum_{m=1}^{I_j'} Y_{mj}\right)$$

which requires arbitrary control of the new jth stratum sample size I_j' and prior knowledge of $P(X_{ik}\in L_j)$, denoted by p_j, for each stratum $j=2,3,\ldots,n+1$.

Figure 5 shows that STRAT reallocates the number of observations per stratum (SA) using the prior knowledge of the strata

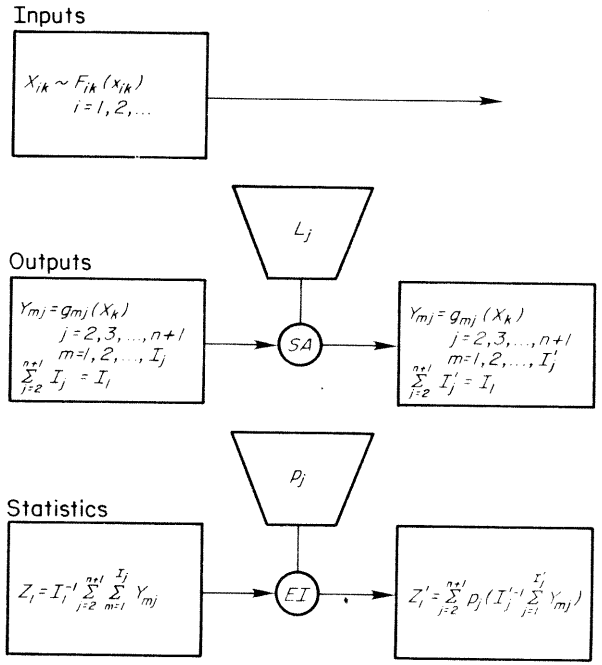

FIGURE 5 Stratified sampling (STRAT).

definitions and then reweights the outputs in the estimator by the ratio p_j/I_j' (EI). Generalization to stratifying on a random vector rather than a scalar random variable does not affect the decomposition. Also, whether or not the new sample allocation yields a variance reduction does not affect the decomposition. Allocation strategies are not discussed here (see for instance, Cochran, 1977), but *proportional allocation* ($I_j'=I_1 p_j$) guarantees

$$\mathrm{Var}(Z_1')\leqq\mathrm{Var}(Z_1)$$

(Rubinstein, 1981). If the I_j are not altered by fixing them in advance, then the VRT is known as poststratifying the sample; see Section 4.5 below.

4.5 Poststratifying the Sample (PSTRAT)

A source of variability in all sampling experiments is that the sample is not representative of the population sampled. Using proportional allocation, STRAT forces a more representative sample by controlling the sampling plan $\{I_2, I_3, \ldots, I_{n+1}\}$. When such control is not possible, sometimes each observation can be weighted on the basis of whether its stratum is under- or over-represented in the sample. This VRT, usually called *poststratified sampling*, we call *poststratifying the sample*, since the sampling plan is not altered. Bratley, Fox and Schrage (1983), Cochran (1977), and Kleijnen (1974) discuss PSTRAT. Wilson and Pritsker (1984a, b) apply PSTRAT in queueing simulation.

Using the above notation for STRAT, consider the PSTRAT estimator

$$Z_1'' = \sum_{j=1}^{n+1} \sum_{m=1}^{I_j} \frac{p_j}{I_j} Y_{mj}$$

where Y_{mj} is the mth observation of Y_1 for which $X_{ik} \in L_j$, and I_j is the number of such observations of the I_1 total. The I_j are outputs (the result of sampling), but we continue to denote them as I_j for convenience.

Provided $I_j > 0$ for $j = 2, 3, \ldots, n+1$, Z_1'' is an unbiased estimator of θ_1. Whereas Z_1 gives each observation weight $1/I_1$, Z_1'' gives weight p_j/I_j. If the observations distribute themselves proportionately ($I_j = p_j I_1$) then this reduces to $1/I_1$. If a stratum is over- or under-represented, p_j/I_j is less or greater than $1/I_1$, respectively. Thus PSTRAT corrects for disproportionate sampling, in contrast to CV, which corrects for shifts in location. Figure 6 shows how PSTRAT combines the strata sample sizes as auxiliary information (AI) with the prior knowledge of the strata probabilities to obtain the new estimator (EI). In our taxonomy PSTRAT is more closely related to CV than to STRAT, since I_j in PSTRAT is random and therefore an output requiring AI, while in STRAT I_j' is a known constant and therefore simply part of the new estimator.

4.6 Conditional Expectations (CE)

The conditional expectations method is often called *conditional*

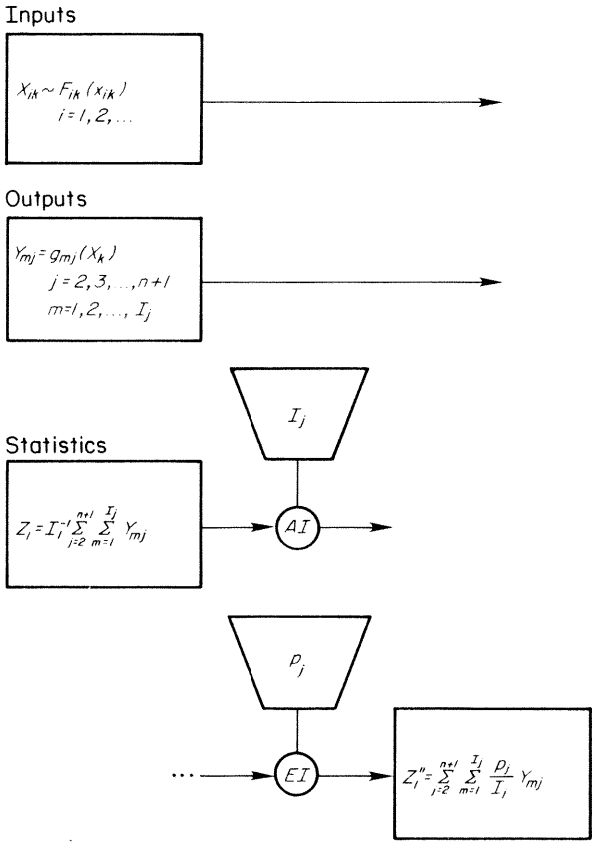

FIGURE 6 Poststratifying the sample (PSTRAT).

Monte Carlo (CMC). However, CMC also refers to a sampling technique developed by Trotter and Tukey (1956) to "use a family of transformations to convert given samples into samples conditioned on a given characteristic" (p. 64). Dubi and Horowitz (1979), Granovsky (1981), and Wilson (1985) discuss CMC in detail. But other than to mention that the original CMC employs a transformation in EA, we do not discuss CMC further here, since the background and detail needed to decompose CMC requires the precise definitions of our taxonomy.

We reserve the term *conditional Monte Carlo* for the original sampling technique. *Conditional Expectations* (CE) is used here as in Law and Kelton (1982), where the expected value of the output of interest is replaced by its known conditional expected value.

Consider estimating θ_1 using a simulation experiment defining

$$Y_{i1} = g_{i1}(X_{(i1)}) \quad i = 1, 2, \ldots, I_1$$

where $E(Y_{i1}) = \theta_1$, with statistic

$$Z_1 = h_1(Y_1) = I_1^{-1} \sum_{i=1}^{I_1} Y_{i1} \tag{7}$$

If there is another output random variable Y_{i2} $i = 1, 2, \ldots, I_2$ such that

$$E(Y_1 \mid Y_{i2} = y_{i2})$$

is known or can be calculated for all realizations y_{i2} of Y_{i2}, where here Y_1 is generic for any of Y_{i1}, then based on the well-known relation

$$\mathrm{Var}(E(Y_1 \mid Y_{i2})) = \mathrm{Var}(Y_1) - E(\mathrm{Var}(Y_1 \mid Y_{i2}))$$

the conditional expectation estimator

$$Z'_1 = h'_1(Y_2) = I_2^{-1} \sum_{i=1}^{I_2} E(Y_1 \mid Y_{i2}) \tag{8}$$

can be used. As shown in Figure 7, based on the (possibly only suspected) prior knowledge that $E(\mathrm{Var}(Y_1 \mid Y_{i2})) > 0$, CE uses Y_{i2} as auxiliary information (AI) in the modified estimator (EI) based on the constants $E(Y_1 \mid y_{i2})$ obtained from prior knowledge.

The estimator (8) is unbiased for θ_1, and if $I_1 = I_2$ and the Y_{i2} are independent then it has no greater variance than (7). However, CE is often employed when $I_2 > I_1$, such as when Y_{i1} are results of "rare events". Clearly the estimator (8) may be based on a vector of outputs, not just a scalar Y_{i2}, but this does not affect its decomposition. Note that Y_{i1} has not been redefined, but rather other outputs (auxiliary information) in the simulation experiment are used.

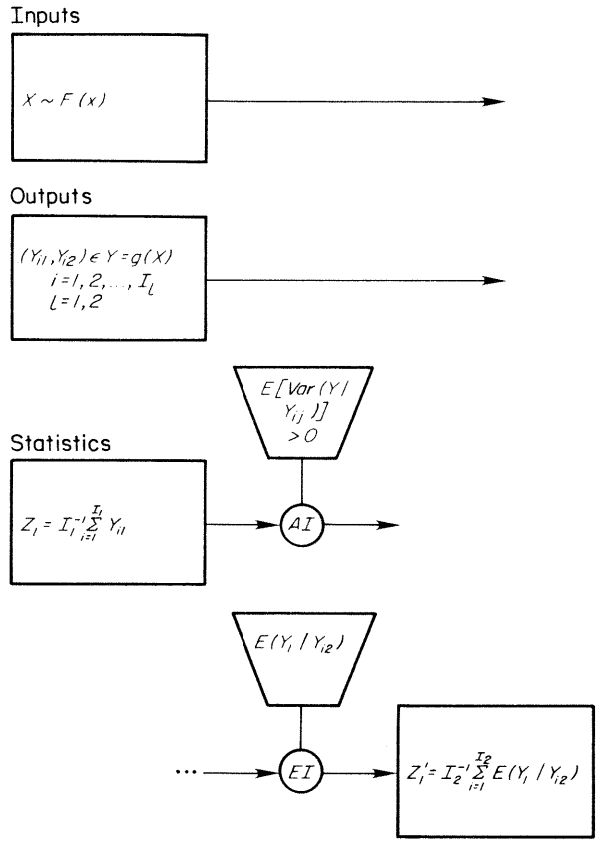

FIGURE 7 Conditional expectations (CE).

4.7 Importance Sampling (IS)

Importance sampling is one of several VRTs that attempt to concentrate sampling in regions of interest, where *interest* may be related to the variance within the region, the likelihood of observations in the region, and/or the magnitude of observations in the region. In contrast to STRAT, which directly allocates sampling effort, IS biases the outputs by altering the probability distributions of the inputs.

IS is a standard technique in Monte Carlo estimation problems; see for instance Hammersley and Handscomb (1964) and Kahn (1956). The technique is used infrequently in systems simulation because the effect of altering the input distribution is often difficult to derive. See Kleijnen (1974) for a general discussion, and Jeruchim (1984) for an example.

A simple version of IS that illustrates the central idea is given here. Consider estimating θ_1 by I_1 observations of Y_{i1}, where $E(Y_{i1}) = \theta_1$, $i = 1, 2, \ldots, I_1$. The crude estimator of θ_1 might be

$$Z_1 = h_1(Y_{(1)}) = I_1^{-1} \sum_{i=1}^{I_1} Y_{i1}$$

Now suppose Y_{i1} is a function of $(X_{(i1)}, X_{ik})$ for some fixed column index k of X. For notational convenience write

$$Y_{i1} = g_{i1}(X_{ik})$$

suppressing the $X_{(i1)}$. Assume that X_{ik} are i.i.d. random variables for all i with identical discrete or continuous marginal pdf $f_{ik}(x_{ik})$. Consider some other pdf f'_{ik} of the same type and having the same support as f_{ik}. If X_{ik} is sampled from f'_{ik} and if $X_{(i1)}$ is sampled from the unaltered conditional distribution of $X_{(i1)}$ given $X_{ik} = x_{ik}$, then

$$Z'_1 = I_1^{-1} \sum_{i=1}^{I_1} Y_{i1} \frac{f_{ik}(X_{ik})}{f'_{ik}(X_{ik})} = I_1^{-1} \sum_{i=1}^{I_1} Y_{i1}\, Y_{i2}$$

is the (unbiased) IS estimator of θ_1. As shown in Figure 8, IS employs the new input distribution f'_{ik} (DR), the new outputs Y_{i2} based on the prior knowledge of the old and new input distributions (EA), and the new statistic (EI) that averages the product of the original outputs Y_{i1} and the auxiliary information Y_{i2} (AI). Generalization to altering the distribution of a random vector, rather than the scalar X_{ik}, does not change the decomposition.

The decomposition also does not depend upon the choice of the new input distribution $f'_{ik}(x_{ik})$, which typically is chosen to be approximately proportional to $|E(Y_{i1}|x_{ik})| f_{ik}(x_{ik})$, where the expectation is over $X_{(i1)}$. This measure of importance is roughly the product of output magnitude $|y_{i1}|$ and input likelihood. See Kahn (1956) for additional details.

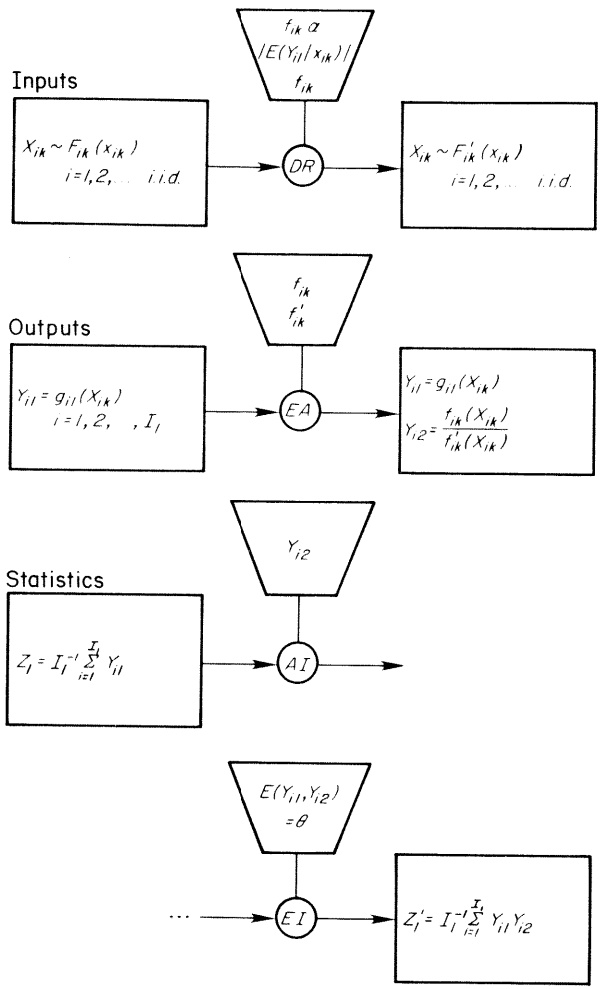

FIGURE 8 Importance sampling (IS).

5. CONCLUSIONS

The examples in this paper illustrate the variance reduction taxonomy developed in Nelson and Schmeiser (1984a, b). The analysis of variance reduction techniques by decomposing them into mem-

bers of the six classes of transformations is central. Other attempts to develop a taxonomy of variance reduction, such as McGrath and Irving (1973) and Wilson (1983a, 1985), were not completely successful, in part because they tried to categorize VRTs. But as we have seen in Section 4, simple partitions do not exist. VRTs overlap in the sense that they are composed of transformations from common classes.

Returning to the question raised in the introduction—Why is a taxonomy of variance reduction useful?—we now discuss the five criteria mentioned in the introduction in light of the taxonomy presented in this paper:

a) The taxonomy reduces the confusion that currently exists regarding the fundamental characteristics of, and relationships among, VRTs. Nelson (1983) listed eighty names of VRTs that have appeared in the literature. Many are different names for almost (if not precisely) the same technique and many fundamentally different techniques have quite similar (or identical) names. These similarities and differences are more apparent when VRTs are decomposed into their basic transformations, as the examples in this paper illustrate.

b) The taxonomy provides a common language for communication among researchers, between researchers and practitioners, and among practitioners. The problem of multiple names for roughly equivalent (where we can now say what roughly equivalent means) VRTs should be less of an issue. The similarities and differences between VRTs are apparent when they are decomposed into their elemental transformations. The existence of classes of transformations makes possible the unification of theoretical results concerning conditions that insure effectiveness of VRTs, results that presently exist in a fragmented fashion in the literature and have been of limited practical use.

Ideally the taxonomy would also be a useful tool for teaching variance reduction to simulators (students and practitioners), but our experience in this area has been disappointing. The reason, and a motivation for this paper, is that learning proceeds most naturally from specific to abstract.

c) The taxonomy provides a different, and we think more effective, approach for practitioners to find variance reduction strategies in simulation experiments. Our experience has been that once a simulation experiment is expressed in terms of the inputs, outputs and statistics, then examining the experiment from the perspective of the six classes of transformations facilitates the identification of appropriate variance reduction strategies. In a variety of real and textbook cases, we have found that thinking in terms of the taxonomy provided a ready-made list of ideas (the six classes), helped identify the relevant prior knowledge, suggested generalizations and modifications of standard techniques, and indirectly provided a sense of when to stop the search for more variance reduction ideas. Of course, it is difficult to determine whether the improved ability to find variance reduction ideas is due to the insight gained from the months spent creating the taxonomy or from the taxonomy itself. Only the experience of others will tell.

d) The taxonomy generates new variance reduction ideas. While our taxonomy is not an "erector set" of components from which VRTs are directly assembled, it does foster discovery of new ideas by expanding rigid definitions of standard VRTs (see CV above, for example). In terms of the six classes of transformations, no VRT is really new, it is just an expansion of existing ideas made possible by a broader perspective. Examples of such ideas that the authors have developed (and have not, to our knowledge, appeared in the literature) include: (i) using nonlinear control variates for estimating probabilities, where standard linear controls may yield infeasible estimates, (ii) using the difference between two analytic results as control variates, and (iii) using distribution replacement to make optimal sample allocation possible (a difficult concept before the two, often confused, ideas were distinctly separated).

e) The taxonomy provides a basis for research on automated variance reduction for general simulation experiments. The underlying motivation for the creation of our taxonomy was the observation that variance reduction, despite its obvious benefit, is seldom used because of the difficulty in learning variance reduction ideas, the difficulty in implementing many of the techniques, and the likelihood of improperly implementing a technique, thereby invalidating the experiment. Any hope of widespread use of variance reduction depends upon the automation of these methods. Automation needs to occur for both the identification of appropriate variance reduction ideas and for the implementation. The long term goal is an interactive system capable of looking at a computer simulation

model and asking good questions about the available prior knowledge (recall the definition of prior knowledge) and suggesting good variance reduction ideas. In conjunction with a particular language such a system would then implement the ideas. But any automated system needs a "world" in which to work. We hope that, just as our taxonomy has provided a world view for us to identify appropriate VRTs directly, the taxonomy will do the same for automated systems.

In discussing these five criteria, we are making an important distinction between the design and the analysis of VRTs. Criteria (c), (d), and (e) above concern design, while the present paper illustrates the analysis aspect (criteria (a) and (b)) by demonstrating how VRTs are decomposed.

We are often asked: "In terms of this taxonomy, what is VRT x?" Our answer is usually to give the decomposition for what we consider to be the most common form of VRT x, as was done in this paper. A better answer would be "it depends". A strength of our taxonomy is that it does not categorize VRTs, but reflects what is actually transformed in the simulation experiment. The decomposition of a VRT depends, for instance, on what the crude experiment was. Also, two techniques that emphasize sampling based on *importance* may differ greatly in how they achieve it (see STRAT and IS, above). In this paper, the decomposition of each VRT should be considered in light of the definition of the VRT presented here; it may not (and probably will not) be the same for every application that goes by the same name.

Tables I and II summarize the decompositions presented in this paper. Any decomposed VRT lies in a cell of one of these or a higher order table. Six of the seven VRTs discussed are shown. The seventh, importance sampling, lies in the fourth order table in cell (DR, EA, EI, AI).

TABLE I
VRT single-component decompositions

DR	DI	EA	SA	EI	AI
	CRN AV				

TABLE II
VRT two-component decompositions

	DR	DI	EA	SA	EI	AI
DR	●					
DI		●				
EA			●			
SA				●	STRAT	
EI					●	CE, CV, PSTRAT
AI						●

We conjecture that if all known VRTs were decomposed and entered into these and higher order tables, there would be empty cells. Why are these cells empty? Do they suggest new VRTs to be discovered, or are there some combinations that are infeasible? Our taxonomy suggests the openings and provides a foundation on which to develop further structure for addressing such issues.

Acknowledgements

The Office of Naval Research provided partial support via contract N00014-79-C-0832. We thank James R. Wilson for both general discussions and specific suggestions that have improved this paper and clarified our thoughts.

References

Bratley, P., Fox, B. L. and Schrage, L. E. (1983). *A Guide to Simulation.* Springer-Verlag, New York.

Cochran, W. G. (1977). *Sampling Techniques.* Wiley, New York.

Dubi, A. and Horowitz, Y. S. (1979). The interpretation of conditional Monte Carlo as a form of importance sampling. *SIAM J. Appl. Math.* **36,** 1, 115–122.

Fishman, G. S. and Huang, B. (1983). Antithetic variates revisited. *Commun. ACM* **26,** 11, 964–971.

Granovsky, B. L. (1981). Optimal formulae of the conditional Monte Carlo. *SIAM J. Alg. Discrete Meth.* **2,** 3, 289–294.

Halton, J. H. and Handscomb, D. C. (1957). A method for increasing the efficiency of Monte Carlo integration. *J. ACM* **4,** 3, 329–340.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods.* Chapman and Hall, London.

Hammersley, J. M. and Mauldon, J. (1956). General principles of antithetic variates. *Proc. Camb. Phil. Soc.* **52,** 476–481.

Hammersley, J. M. and Morton, K. W. (1956). A new Monte Carlo technique: Antithetic variates. *Proc. Camb. Phil. Soc.* **52,** 449–475.

Handscomb, D. C. (1958). Proof of the antithetic-variates theorem for $n>2$. *Proc. Camb. Phil. Soc.* **54,** 2, 300–301.

Isaki, C. T. (1983). Variance estimation using auxiliary information. *J. Am. Statist. Assoc.* **78,** 381, 117–123.

Jeruchim, M. C. (1984). On the application of importance sampling to the simulation of digital satellite and multihop links. *IEEE Trans. Communications* **32,** 10, 1082–1088.

Kahn, H. (1956). Use of different Monte Carlo sampling techniques. In *Symposium on Monte Carlo Methods.* (Ed. H. Meyer), Wiley, New York, 146–190.

Kleijnen, J. (1974). *Statistical Techniques in Simulation, Part I.* Marcel Dekker, New York.

Law, A. M. and Kelton, W. D. (1982). *Simulation Modeling and Analysis.* McGraw-Hill Book Co., New York.

McGrath, E. J. and Irving, D. C. (1973). Techniques for efficient Monte Carlo simulation, Vol. III, Variance Reduction. ORNL Report, SAI-72-509-LJ.

Morton, K. W. (1957). A generalization of the antithetic variate technique for evaluating integrals. *J. Math. Physics* **36,** 3, 289–293.

Nelson, B. L. (1983). Variance reduction in simulation experiments: A mathematical-statistical framework. Unpublished Ph.D. dissertation, School of Industrial Engineering, Purdue University.

Nelson, B. L. and Schmeiser, B. W. (1984a). A mathematical-statistical framework for variance reduction, Part I: Simulation experiments. Research Memorandum 84-4, School of Industrial Engineering, Purdue University.

Nelson, B. L. and Schmeiser, B. W. (1984b). A mathematical-statistical framework for variance reduction, Part II: Classes of transformations. Research Memorandum 84-5, School of Industrial Engineering, Purdue University.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method.* Wiley, New York.

Trotter, H. F. and Tukey, J. W. (1956). Conditional Monte Carlo for normal samples. In *Symposium on Monte Carlo Methods.* (Ed. H. Meyer), 64–79, Wiley, New York.

Whitt, W. (1976). Bivariate distributions with given marginals. *Ann Statist.* **4,** 1280–1289.

Wilson, J. R. (1979). Proof of the antithetic variate theorem for unbounded functions. *Proc. Camb. Phil. Soc.* **86,** 477–479.

Wilson, J. R. (1983a). Variance reduction: The current state. *Math. Comput. Simul.* **XXV,** 55–59.

Wilson, J. R. (1983b). Antithetic sampling with multivariate inputs. *American J. Math. and Mgt. Sci.* **3,** 2, 121–144.

Wilson, J. R. (1984). Variance reduction techniques for digital simulation. *American J. Math. and Mgt. Sci.,* **4,** (3/4), 277–312.

Wilson, J. R. and Pritsker, A. A. B. (1984a). Variance reduction in queueing simulation using generalized concomitant variables. *J. Statist. Comput. Simul.* **19,** 129–153.

Wilson, J. R. and Pritsker, A. A. B. (1984b). Experimental evaluation of variance reduction techniques for queueing simulation using generalized concomitant variables. *Mgt. Sci.* **30,** 12, 1459–1472.