Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

CAVEATS FOR SIMULATION MODELING IN SUPPORT OF DECISION MAKING

Christopher M. Hill

Center for Army Analysis
6001 Goethals Road
Suite 102
Fort Belvoir, VA 22060-5230, U.S.A.

Linda C. Malone

Industrial Engineering and Management Systems
University of Central Florida
4000 Central Florida Boulevard
Orlando, FL 32816-2450, U.S.A.

ABSTRACT

Simulation modeling is a tool commonly used in support of intelligent decision making by senior managers, particularly for extremely complex problems. This article uses an example from the United States Army Recruiting Command to illustrate some of the statistical pitfalls an analyst may encounter when using simulation modeling. These pitfalls include conflicting results, both due to different modeling approaches and choice of input distributions, and incorrect interpretation of the simulation experimental results. The paper also provides implications for analysts who encounter these situations. The analyst who uses simulation in support of senior decision-makers must understand simulation's capabilities, limitations, and statistical underpinnings. Failing to do so can result in decisions based on incorrect information. Analysts can guard against these pitfalls through careful consideration of statistics, preparation, and communication.

1 INTRODUCTION

The United States Army provides a good example for how analytical support helps senior management make intelligent decisions. The Army relies on its pool of operations researchers to produce analyses and other analytic products to underpin decisions by leaders and managers at all levels of the Department of Defense (DOD), and to enable solutions to varied and complex strategic, operational, tactical and managerial issues (DA PAM 600-3, 1998). Army leaders, like their civilian counterparts, must make myriad, complex business types of decisions concerning the daily institutional operations. However, the Army senior management must also make decisions required by the operational forces, which are either deployed or preparing for combat operations. Discrete event and other simulation models typically support these functions because of their ability to reasonably depict extremely complex systems. In many cases, a simulation model may be the only feasible alternative for the military analyst. The military analyst who uses simulation in support of senior decision makers, as well as their civilian counterparts, must understand its capabilities and limitations, as well as its statistical foundations in order to avoid misleading results.

In the Army, a large percentage of the simulation work routinely falls in the combat modeling application. However, analysis of the military's daily institutional operations is just as relevant for simulation models. One example of an organization that is increasing its use of simulation models for its business decisions is the United States Army Recruiting Command (USAREC). Simulation is a tool with powerful potential, but failure to understand the statistics underlying applications can lead to erroneous inferences. This paper will explore some of the more sensitive areas in terms of causes of the problems and implications for analysts. A contrived simulation example from USAREC using ARENA software will provide a detailed example of each point.

2 SIMULATION'S POTENTIAL PITFALLS

Simulation studies can be powerful tools if developed with a solid statistical foundation and with correct analysis and interpretation of results. Inadvertent mistakes can occur because many of the real world systems under study and their associated simulation models are very complex. Even with correctly built and executed models, it is still easy to apply statistics incorrectly and reach faulty conclusions. The two potential pitfalls we will discuss are conflicting results from different simulation models and incorrect interpretation of the output measures.

2.1 Conflicting Results

The first area of concern is the potential for conflicting results. It is conceivable and even likely that two well-qualified and knowledgeable analysts who are modeling the same data set can achieve different models for the system under study. More disturbing, it is also possible for the same two modelers to reach different conclusions when modeling the same scenarios from the same data from a common system under study. This is possible because of differing statistical analysis of the data at the front end of the study that yields different models and because of different choices of input distributions.

2.1.1 Model Differences

There are several reasons why two modelers may arrive at different simulation models. The process of a simulation study is the representation of a real system with stochastic and random data in a synthetic world for the purposes of studying some specific aspect of the real world system (Vincent, 1998). The data by its very nature will vary to some extent. We assume for the purposes of this article that the data is available, and is observed rather than collected by a deliberate experiment. These assumptions are consistent with the recruiting example. The first step of input modeling is generally to assess the data in terms of its relationships and independence, and to see if assumptions concerning independence and identical distribution (iid) are valid. The next step is to observe the data more closely through plots, like histograms, and summary level statistics. These steps will help determine at what level and grouping to model the data. This is the where the analytical water begins to get murky. Much of data analysis is art rather than science. With extremely complex data, the relationships may be difficult to determine, and could lead to different although reasonable modeling approaches. This first step can be the source of great divergence among analysts.

An example can best describe this condition. Suppose the United States Army is considering the addition of E-4 Corporal (CPL) recruiters to its force. Before 1999, the recruiting force in the Army did not allow soldiers of the lower enlisted ranks to participate in recruiting operations other than short-term supplementation through the Hometown Recruiter Assistance Program (HRAP). Although this decision-making scenario is historically accurate, our specific example is entirely contrived. The genesis of the decision is that the lower grade enlisted may improve the recruiters ability to communicate and relate with youth while relieving the force of some of its large requirements for mid-grade non-commissioned officers. This type of decision serves as an excellent example because it is well suited for the use of simulation in support of senior decision-making. No recruiters in the rank of CPL existed in the system, so there was no data available to model their behavior. In addition, the decision to add CPL recruiters would eventually need to be made at the four-star level, so significant numbers and levels of analysts would become involved.

Suppose the commander of USAREC is in the process of forming his position on CPL recruiters, and he has asked his Program and Evaluation Directorate (PAE) for supporting analysis and a recommendation. The director assigns the task of modeling the scenario to two divisions—Research and Plans (R&P) and the Strategic Plans Office (SPO). These two offices will provide different perspectives on the issue. The director feels the most likely point in the recruiting process where a CPL can influence the process is after a recruiter has made an appointment with a prospect. In this process, the prospect arrives into the station; the recruiter will establish rapport with the individual and conduct a basic qualification assessment of education level, mental ability, and moral standards. After these steps if the recruiter feels the prospect is qualified and available, he will begin a detailed sales pitch with specific individual features oriented to the prospect. If the prospect remains interested, the recruiter closes the deal by completing an application and making an appointment at the Military Entrance Processing Station (MEPS) where detailed qualification and contracting is completed. The real-world process is actually very complicated with random attrition behavior present at all phases in the process and walk-in traffic arriving in addition to appointments, but we will only consider this simplified flow to allow focus on the statistical issues. In addition, there could be multiple measures of effectiveness for this study, but we will only focus on a prospect's mean time in the system.

The analyst from R&P considers the problem and determines the best way to model the data is by using a logic model shown below in Figure 1. He believes the best approach is to merge the rapport and basic qualification phases because of their frequent overlap during execution. He also believes in pairing the closing and application phases, since many recruiters will work on these phases simultaneously. After lengthy discussion with subject matter experts, he also believes the CPL may influence each of these processes at different levels.

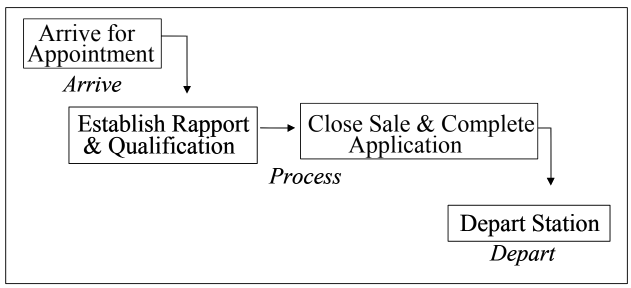

Figure 1: Research and Plans Model

The analyst obtains data on a typical recruiting station to build the base model, using Arena simulation software. The analyst uses an arrival rate and two processing distributions. The analyst verifies and validates his model and then runs a series of 30 replications for comparison with the SPO office. The results of the simulation experiment for the mean time in station are $\mu rp = 72.88$ and $\sigma rp = 7.70$.

The analyst from SPO takes a different perspective. The SPO already has a strategic level simulation model, also using Arena simulation software, which he feels will be sufficient for analysis of the CPL problem with slight modification. Figure 2 shows the logic flow of his model. The SPO analyst considers one processing node that combines all four phases of the recruiter-prospect interaction. The SPO analyst considers the same data set and information obtained and used by the R&P analyst for his simulation model. The difference is that the SPO analyst The SPO analyst considers the same data set and information obtained and used by the R&P analyst for his simulation model. The difference is that the SPO analyst combines the process data into one set and builds a single process input distribution for his model. He then conducts verification, validation, and runs 30 replications. The results of his model for mean time in station have $\mu spo = 78.10$ and $\sigma spo = 9.67$.

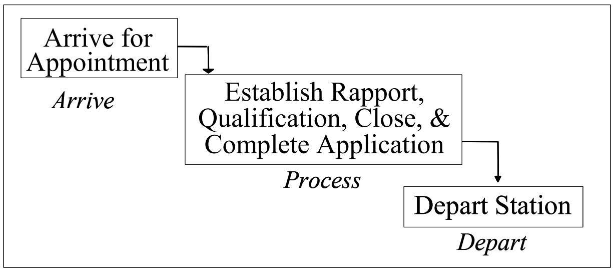

Figure 2: Strategic Plans Office Model

Comparison of the mean time in the station for the two models shows statistically significantly different results. Using a two-sample t-test, without an equal variance assumption, the null hypothesis is $\mu rp = \mu spo$. The test statistic for the difference in the two simulation scenarios is –5.22, which yields a p-value of 0.02. Since the p-value is less than our α value of 0.05, the result is rejection of the null hypothesis, and thus $\mu rp \neq \mu spo$.

Evidence exists substantiating a significant difference in the two modeler's results. The two analysts reach different conclusions although both are well qualified, use logically accurate models, and model the same data at different levels. What does this difference imply for the director of PAE? How does it affect the development of a position on the CPL issue?

Resolution of such analytical conflict needs as much consideration as the statistical choices in simulation modeling. In a conflict such as this where two differing approaches yield different results the key is to model at the level that most efficiently and effectively represents the real system under consideration. The decision-maker must be aware of the differing assumptions of the approaches and how they impact modeling of the proposed changes. The SPO model considers the system at a more aggregated level, which will provide more efficiency but may sacrifice resolution. If the eventual impact of a CPL recruiter occurs at a level that is more disaggregated than in the SPO's single process, it will not be possible to accurately model the effects of change; in this case, the R&P model may be a better choice. If the differences are not easy to determine, or if the impact is not known, a technique employed at Recruiting Command is the "murder board." In this technique, the opposing analysts would present positions to an audience of analysts within the organization that would challenge the assumptions, present issues, and brainstorm on possibilities. These sessions frequently help identify issues that lead to conflict resolution.

2.1.2 Input Distributions

The next area where the modelers may diverge is in selection of input distributions. After understanding the data set, the modeler then seeks to find appropriate sets of input distributions. The modeler will analyze a hypothesized candidate distribution or group of distributions for goodness of fit with the data, and he will attempt to find a reasonable representation (Vincent, 1998). The definition of reasonable will certainly vary between analysts. One can accomplish this through manual statistical calculation, or with software designed to automate and simplify the process. As a modeler attempts to fit a distribution, several alternatives are possible. The first, and probably least likely, is that he finds a candidate theoretical distribution that is the only one that meets appropriate goodness of fit criteria. The second is that there are several distributions that meet the modelers hypothesis, and each results in good fitness scores. The last is that the modeler may find that none of the common theoretical distributions will work. In the last two cases, the method by which the analyst resolves his dilemma will vary. To resolve the problem the analyst should assume the measure of effectiveness of the study is very sensitive to his choice, use prior experience, apply relevant applicable theory, and use the data (Vincent, 1998). Again, this is an area where art may override science, so results among analysts can vary even though the logical structure of the models are identical. In any case, the analyst also faces the choice of an empirical distribution versus a theoretical distribution. This option creates another dimension of input modeling possibilities and trade-offs (Kelton, et al., 1990). Research has even shown that output measures such as mean wait in queue are sensitive to the choice of input distributions (Gross, 1999). As a result, modelers who decide to use different distributions, even in very similarly constructed models, may reach different conclusions. The procedure becomes even more difficult in the absence of data or when simple models fail (Nelson, et. al., 1995 and Schmeiser, 1999).

For another specific example, consider the analyst from the R&P division mentioned earlier. When building the simulation model with two distinct processes, the analyst found several potential input distributions. Using Arena's input analyzer, the set of data yielded four candidate distributions available for input distributions of the Rapport and Qualification Process and six candidate input distributions for the Close and Application phase as shown in Table 1. These choices did not consider using empirical or more complex distributions, such as a Johnson. The set of potential distributions could very easily be larger.

Table 1: Input Distributions

R&Q Distribution Fit Table Process 1				
Distribution	Parameters	Square Error	Chi-Square P-value	KS P-value
Beta	21+27*Beta (2.06. 1.77)	0.113	0.231	>0.15
Normal	Norm (35.5, 6.08)	0.008	0.151	>0.15
Triangular	Tria (21, 36.2, 48)	0.007	0.355	>0.15
Weibull	21+Weib (16.1, 2.41)	0.011	0.226	>0.15
C&A Distribution Fit Table Process 2				
Distribution	Parameters	Square Error	Chi-Square P-value	KS P-value
Beta	9+71*Beta (1.49, 2.14)	0.009	0.38	>0.15
Erlang	9+Erla (14.6, 2)	0.019	0.07	>0.15
Gamma	9+Gamm (16.3, 1.8)	0.021	0.03	0.09
Normal	Norm (38.2, 16.1)	0.005	0.56	>0.15
Triangular	Tria (9, 40.1, 80)	0.015	0.1	0.056
Weibull	9+Weib (32.1, 1.67)	0.011	0.27	>0.15

Consider the impact of changing distributions on the measure of effectiveness of mean time in the station. The first model uses the triangular distribution for the rapport and qualification process and the beta distribution for the close and application process. The results of 30 replications of the model yield $\mu a = 72.88$ and $\sigma a = 7.70$ for our measure of effectiveness. Now, let us change the distribution for rapport and qualification to the beta and the close and qualification to a triangular. This change results in $\mu b = 77.98$ and $\sigma b = 6.67$. A two-sample t-test compares means with a null hypothesis of $\mu a = \mu b$. The test statistic is –5.10 with a p-value of 0.00. Since the p-value is less than our α value of 0.05, we reject the null hypothesis and determine there is a difference in output from the two approaches. In this case, the choice of input distributions significantly affects the output of the model, even though all choices are valid. What does this sensitivity tell the analyst as he is making a recommended position for the director or the senior level decision-maker?

Specific resolution of the choice of input distributions is not a straightforward issue. One way to resolve problems is to consider historic approaches of other researchers modeling similar systems. The analyst must also spend a great deal of time studying the data and considering the impact of various distribution choices, particularly with respect to the tail ends of the distribution. The analyst must also discuss the issue with experts on the real system to determine any characteristics that may influence the choice. Once the choices are narrowed to a small group, the analyst can also conduct sensitivity analysis to learn more about specific impacts of changes in input distributions on output measures and general model performance. Sensitivity analysis may reveal problems with data as well. If the sensitivity of the model is isolated to a particular input, the analyst should consider another data collection if possible. However, this approach is not always feasible. Even in cases when the simulation model remains sensitive to a particular input, the sensitivity analysis is crucial to providing information about the critical aspects of the simulation model.

The situation of conflicting results described above is nothing new to simulation modelers. However, it is important to contemplate the impacts of the potential differences before presenting results on a high priority project to senior managers. Since different analysts, staffs, or agencies working on the same problem may reach differing conclusions, the key to conflict resolution is early and frequent communication about modeling approaches and assumptions. The true benefit of the simulation study is the information it provides about the system, not necessarily a precise output from a model. In addition, the analyst must always strive for robust results that provide useful information in a wide variety of conditions. It is important to discover, compare, and analyze differences in models or approaches before a formal presentation. If the groups are "allies" on the issue, like different members of a single staff office in our example, this step may be easier than if the groups are in different camps.

Confusing the decision maker can be a major problem, and this is particularly true when the decision maker is not an analyst. It is difficult to explain in layman's terms the causes for the differences without preparation and careful thought. Otherwise, the conversation may slip into "statistics-speak" and you will lose the staff, audience, decision maker, and potentially your own credibility.

The next issue is making sure the decision-maker understands the likelihood of different conclusions if he is taking the results to an outside agency, headquarters, or higher-echelon organization. It seems as if the number of analysts investigating a problem increases proportionally with the level of the brief and impact of the decision. Differing conclusions by these groups are almost inevitable. The boss must fully understand the relationships and information learned from a simulation study, not just the final output numbers. If he understands this, he will be prepared to confront differing positions and the study will be successful. Information that supports intelligent decisions, regardless of the specific outcome determines success.

2.2 What Do the Output Measures Really Mean?

The nature of a stochastic simulation model is that random numbers drive the model through various input distributions. As a result, the output numbers are random as well. This has several implications, the least of which is that the analyst must use caution with the statistics used in analysis. However, more importantly, it is easy for the model's random numbers to fool you. These random numbers will have a variance associated with them that directly affect the output measures (Kelton, 1997). This variance can also result in very broad half-widths for confidence intervals and make interpretation difficult. The estimates could differ greatly from the true characteristics of the model and lead the analyst to incorrect conclusions.

In addition, the interpretation of a confidence interval is also easily misunderstood. The correct interpretation of a 90 percent confidence interval is "90 percent of the time the confidence interval formed between the upper and lower bounds of the interval (*a* and *b*) covers μ." An incorrect interpretation is "I am 90 percent confident that μ lies between *a* and *b*." In our simulation example this means that if we ran 100 replications of the R&P or SPO experiments, the μ will be included in 90 of the experiment's confidence intervals, not that there is a 90 percent chance it is included in each of the 100 experimental outputs (Law & Kelton, 2000).

Multiple measures of performance exacerbate the variance problem And there are implications from the Bonferroni inequality. Suppose that Is is a 100(1-α) percent confidence interval for the measure of performance μs. The probability that all *k* confidence intervals simultaneously contain their respective true measure satisfies $P(\mu s \in Is$, for all s=1, 2, …, $k) \geq 1-\sum_{s=1}^{k}\alpha_s$ (Law & Kelton, 2000). It means that if we had five measures of effectiveness with 90 percent confidence intervals for each, the probability that each of the five contain the true measure of performance is an overall confidence of 50 percent rather than 90 percent. This scenario can add even more confusion to the interpretation and explanation of the results. Illustration of multiple measures of effectiveness through our recruiting example is addition of mean number of applicants, queue time in the station, attrition rate, and time of departure of last prospect to the current single measure of effectives of the mean time in the station.

Since the half-widths of the confidence intervals can be broad, conclusions that are meaningful and easily interpreted are difficult to achieve. One way to counter this effect is to conduct larger numbers of simulation runs. However, this approach may not be feasible or effective in all cases. If not, there are several variance reduction techniques available to the simulation practitioner to counter the broad confidence interval problem. These techniques include common random numbers, antithetic variables, control variates, indirect estimation, and conditioning techniques. Variance reduction techniques induce correlation and dependency in an attempt to achieve more narrow half-widths for the confidence intervals. The aim of each technique is to improve statistical efficiency based on improved precision (Law & Kelton, 2000).

Consider our example of the SPO analyst model of the single recruiter process that includes establishing rapport, qualification, closing the sale, and completing an application. The SPO analyst ran his base model in Arena, and then created a second simulation model for comparison of the potential improvement by adding a CPL recruiter. The SPO analyst used a Beta distribution (44+69*BETA(1.38, 1.82, 5)) to model the original system and obtained $\mu base = 78.9$ and $\sigma base = 16.6$ for the mean time in the station after 15 replications of the model. Based on examination of HRAP data and SME input, the analyst built a second model reflecting the improvement of the station with the CPL recruiter. The data showed potential improvements in the range of 10 to 15 percent in the process by adding a CPL recruiter. The improvement data set yielded a triangular input distribution (TRIA(55, 56.3, 94)). The SPO analyst ran 15 independent replications for the improved model and it yielded mean time in the station with $\mu cplindep = 70.43$ and $\sigma cplindep = 3.83$. The analyst compared these results to the base model with a two-sample t-test. The null hypothesis was $\mu base = \mu cplindep$. The test statistic is 1.92 with a p-value of 0.07. Since the p-value is greater than our α of 0.05, we cannot reject the null hypothesis. This indicates there is not enough evidence to suggest a change from adding the CPL recruiter. These results do not please the analyst, so he decides to use common random number seeds to see if there is a difference in the results. In this case, the parameters of the models do not change at all; the only difference is the modeler ensures the second simulation model is using the same string of random numbers to feed the distributions as the first model. The statistics from the first models do not change: $\mu base = 78.9$ and $\sigma base = 16.6$. The results from the CPL model with common random numbers is $\mu cpldep = 68.37$ and $\sigma cpldep = 5.43$. A paired t-test compares the results from both models with common random numbers, since the outputs are now dependent. The null hypothesis is $\mu base = \mu cpldep$. The comparison yields a test statistic of 3.60 with a p-value of

0.003. Since the p-value is less than our α of 0.05, we reject the null hypothesis. The common random number technique seems to improve precision enough to suggest a difference by adding the CPL recruiters. However, is this difference practically significant? Would you be comfortable recommending a position that advocates adding CPL recruiters and improved performance knowing that the results did not hold up in a scenario where simulation runs were independent, but did when they were dependent?

Resolution of the issue of common random numbers depends on the scenario the analyst faces. If possible, using the exact same simulation model, to include random numbers, when comparing a current system with a new proposal should be employed. This approach shows that nothing in the model changes with the exception of the modifications to the system for which the study was designed. In this case any difference between the two models will likely be due to the changes in the system and not a modeling issue. However, variance reduction techniques are not guaranteed to work on all models, and the impact of the same random numbers may affect two models differently (Law and Kelton, 2000). As a result, the last and most important factor is to remember the relationship between statistical and practical significance. The ability to obtain a "statistically significant" result should not be the sole focus of simulation studies. Consider significance from both a statistical and practical vantage, and weigh the level of difference between alternative scenarios that engenders a practical difference together with choice of analytical techniques and variance reduction methods.

3 IMPLICATIONS

Several statistical pitfalls of simulation modeling have been presented along with specific, simplified examples. Each example contained specific implications for each of the potential pitfalls. These are easily summarized into four groups: careful consideration of statistics, preparation, resolution, and communication. Simulation studies are complex endeavors, and as the real world systems under consideration become more complex, the opportunity to build misleading simulations and to incorrectly apply statistics increases. Analysts must ensure that they have a solid statistical background and carefully consider the statistical implications of each step of the study. In addition, the study should include a sensitivity analysis component to provide specific detail on the portions of the model that may have the greatest potential for influencing results.

Analysts must consider these relationships and conduct detailed preparation before embarking on a study in support of senior management. The analyst must base the modeling approach on a solid theoretical foundation, in terms of both the dynamics of the system under consideration and the statistical tools used in the course of analysis. This sound beginning will allow an analyst to consider alternative analysis options, and will result in a robust study. The analyst should be able to explain his analysis as compared to competing models, and more importantly to provide good information to senior decision-makers.

Resolution of differing results in multiple approaches or models demands the same level of careful consideration as the other statistical aspects. Model at the level, which describes the system under consideration at the best efficiency and effectiveness. Consider sensitivity analysis and the information it provides. Use a critical review process; it can illuminate differences in assumptions as well as their implications. Finally, always be aware of agendas accompanying different modeling approaches. Remember the larger goal of obtaining information about the system under study without putting too much emphasis on the details. Remember the big picture and the risk accompanying all options.

Communication is another key to success both with other analysts and with decision-makers. Since different, competing analyses can easily occur, communication with other analysts is extremely important to facilitate explanation and examination of each approach. Other analysts' results can blind-side you based on a lack of communication. Allowing a senior manager not fluent in statistics to discover competing scenarios by chance will result in disaster. Complex and surprising analyses can be difficult to convey to simulation novices, especially without preparation. Those unfamiliar with statistics and simulation may perceive this as the use of "smoke and mirrors" to get to a predetermined answer. The objectives of the study should always be to consider the whole process in an unbiased manner and base the study on a sound foundation based on expert level, experiential, and theoretical knowledge from both a system and simulation model point of view.

Conflict resolution is another significant aspect of simulation studies. In each specific scenario, the example concluded with questions posed to the analysts or leaders. The methods by which conflicts are resolved need as much consideration as do the statistical choices upon which the analysis is based.

4 CONCLUSION

Simulation tools are crucial in aiding the analyst supporting intelligent decisions by senior managers. The use of these tools is proliferating in the United States Army in applications for daily operations support and is not limited just to the combat modeling environment. These tools are just as relevant in the business sector. The analyst who uses simulation in support of senior decision-makers must understand its capabilities, limitations, and statistical underpinnings. Failing to do so can result in poor interpretation and/or use of output measures. Worse yet, it could result in a decision maker resorting to a "seat-of-the-pants" or "that's how its always been done" decision which can be infinitely worse than intelligent use of simulation even with pitfalls. Analysts can guard against these pitfalls through careful consideration of statistics, preparation, and communication.

ACKNOWLEDGMENTS

The authors would like to thank Mr. P. J. Diss of the United States Army Recruiting Command's Strategic Plans Office for his input and subject matter expertise. We would also like to acknowledge Colonel Greg Parlier, a former Director of the Program Analysis and Evaluation Directorate, United States Army Recruiting Command, for his critical review of the article.

APPENDIX A: MODELING LOG

The purpose of this appendix is to provide the modeling methodology used to build the contrived simulation model examples and the desired statistical properties provided in the article. The intent of the article was to develop trivial simulation models that would realistically illustrate the potential statistical pitfalls of using simulation modeling. Our general approach was to keep most of the factors in the various models constant and only change the specific parameters necessary to illustrate each point. Each simulation model used the same arrival scheme: scheduled arrivals beginning at 0800 in the morning with seven total applicants staggered by a constant 80-minute interval. Each simulation used only one recruiter in the process nodes. The measure of effectiveness for the simulated study was mean time in the system for the applicants.

Our modeling process differed slightly from the manner in which we presented the results in the article. Rather than start with the comparison of differing models, we started with using different input distributions in the same model. We then looked at separate models next and finished with the variance reduction approach.

The modeling process began by using subject matter expert information about the process to create distributions that realistically reflected the process. Based on their information, we created a distribution for Rapport and Qualification (RQ) that was normal with mean 36.5 and standard deviation of 6.67; Close and Application (CA) had normal with mean 37.5 and standard deviation of 15. These base distributions were used in Minitab to generated random numbers from each of them that would serve as input data from which to model input distributions. We conducted these steps to simulate the type of data we might have collected in a real study of recruiting stations. The statistics for these sample distributions was mean 35.51 and standard deviation of 6.124 for RQ and mean 38.2 and standard deviation of 16.23 for CA. These sets of data then went to the Arena input analyzer and fit distributions for each of them. The RQ fit four candidate distributions and the CA fit six candidate distributions. Table 1 of the article shows these results. To fit a distribution for the combined process, we added the two 70 observation samples to create a combination sample. Again, the Arena input analyzer fit distributions and was able to fit six candidate distributions. The result of this step was creation of data used for input modeling and a set of candidate input distributions for each process.

The next step in the process was creating a simulation model of the system. The intent was to create a simple model illustrating the statistical issues, so we ignored many of the real-world complications, like attrition. Figure 1 of the article shows the diagram representing this model. The first model was a system representation with two processes and a very simple arrival scheme. We conducted 30 replications of this model to use as a base output result. This model had a TRIA (21, 36.2, 48) in the Rapport and Qualification process module and a 9+71*BETA (1.49, 2.14) in the close and application.

The next step was to use different input distributions in the two processes of the model to obtain a statistically different output. We built a second model that changed the distributions to 21+27*BETA(2.06, 1.77) in the Rapport and Qualification process module and a TRIA (9, 40.1, 80) in the close and application. We also ran this model at 30 replications. We ensured the models ran with independent replications and used Minitab to conduct a comparison of means. Results from a two-sample t-test concluded the means are not equal between the two models. We describe these results in detail in the input distribution section of the article.

After obtaining differing results for alternate input distributions, we desired to model the system with a different approach to try and get statistically differing results. The model representing the alternative approach combined the two processes of the first model. This model used the 44+69*BETA (1.38, 1.82) input distribution for the single process and ran for thirty replications. Again, we took the data to Minitab to compare means. The results show that the means are different as explained in detail in the Model Differences section of the article. The intent of this step was demonstration that you can have two statistically different results using two different and reasonable modeling approaches from the same data set.

The final and most complex step was demonstration of the variance reduction by use of common random issue. Our methodology began by trying to create a set of data that represented a change or improvement by substituting the CPL recruiter with the more senior recruiters in the recruiting station. In other words, the model reflecting the CPL recruiter would still have only one recruiter in the process node. The same basic simulation model would run the change. The only changes would be the process input distribution reflecting the improvement, and the random number seed to reflect either independent or correlated runs. The objective of this step was that the change in the model would not show a significant difference in the output when using independent runs, but would show a significant difference using common random numbers.

To obtain the hypothetical change or improvement in the process by changing the recruiter from a senior NCO to a CPL recruiter we took the original 30 data points used to build the input distribution for the single process model and randomly generated improvements between 10 and 15 percent. This new data set was taken to Arena to fit input distributions, and we were only able to fit two distributions in Arena's input analyzer, 55+39*BETA (1.28, 2.44) and TRIA (55, 56.3, 54). For the first series of simulation results, we set the random number seeds in the Arena software to ensure independent trials. The improvement model with the Beta distribution ran for 15 replications and analysis of the results in Minitab showed no difference in the means with a two-sample t-test. The first step of our variance reduction example was complete—results with independent runs.

The next step was to keep the same distributions in the processes and run a correlated simulation experiment, with the hopes of achieving statistically different results. We set the Arena software to ensure use of the same random number seeds, which provided correlated runs. Analysis of 15 replications in Minitab with a paired t-test yielded a statistical difference. The details of the variance reduction using the common random number technique is explained in the "What do the Output Measures Really Mean?" section of the article. This step achieved all results needed for the examples in the article.

REFERENCES

Gross, D. 1999. Sensitivity of Output Performance Measures To Input Distribution Shape in Modeling Queues—3: Real Data Scenario. In Farrington, P.A., Nembhard, H. B., Sturrock, D. T., & Evans, G. W. (Eds.), *Proceedings of the 1999 Winter Simulation Conference*, 452-457.

Kelton, D. W., Fox, B. L., Johnson, M. E., Law, A. M., Schmeiser, B. W., and Wilson, J. R. 1990. Alternative Approaches for Specifying Input Distributions and Processes (Panel). In Balci, O., Sadowski, R. P., and Nance, R. E. (Eds.), *Proceedings of the 1990 Winter Simulation Conference*, 382-386.

Kelton, D. W. 1997. Statistical Analysis of Simulation Output. In Andrado'ttir, S, Healy, K. J., Withers, D. H., and Nelson, B. L., (Eds.), *Proceedings of the 1997 Winter Simulation Conference*, 23-30.

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

Law, A. M. & McComas, M. G. 1999. Expertfit: Total Support for Simulation Input Modeling. In Farrington, P.A., Nembhard, H. B., Sturrock, D. T., & Evans, G. W. (Eds.), *Proceedings of the 1999 Winter Simulation Conference*, 452-457.

Nelson, B. L., Ware, P., Cario, M. C., Harris, C. A., Jamison, S. A., Miller, J. O., Steinbugl, J., & Yang, J. 1995. Input Modeling When Simple Models Fail. In Alexopoulos, C., Kang, K., Lilegdon, W. R., & Goldsman, D. (Eds.), *Proceedings of the 1995 Winter Simulation Conference*, 93-100.

Schmeiser, B. 1999. Advanced Input Modeling for Simulation Experimentation. In Farrington, P.A., Nembhard, H. B., Sturrock, D. T., & Evans, G. W. (Eds.), *Proceedings of the 1999 Winter Simulation Conference*, 452-457.

United States Army Publishing Agency. 1 October 1998. Commissioned Officer Development and Career Management. Department of the Army Pamphlet 600-3, 33.1.A. Washington, Headquarters, Department of the Army.

United States Army Recruiting Command. 30 November 1998. United States Army Recruiting Command Regulation 350-6, Training, Recruiter Production Management System. Headquarters United States Army Recruiting Command, Fort Knox, KY.

Vincent, S. 1998. Input Data Analysis. In J. Banks (Ed.), *Handbook of Simulation, Principles, Methodology, Advances, Applications, and Practice*, 55-93. New York: John Wiley & Sons, Inc.

AUTHOR BIOGRAPHIES

CHRISTOPHER M. HILL is an analyst at the Center for Army Analysis at Fort Belvoir, Virginia. He is a Lieutenant Colonel in the United States Army. His research interests include large data analysis, simulation input modeling, and simulation output analysis. His e-mail is <Christopher.Hill@caa.army.mil>.

LINDA C. MALONE is a Professor in the Industrial Engineering and Management Systems Department at the University of Central Florida. Her research interests include experimental design, large data analysis, and simulation output analysis. You can reach her by email at <lmalone@mail.ucf.edu>.