2009

Efficient Experimental Design Tools for Exploring Large Simulation Models

Oh, Regime

Computational and Mathematical Organization Theory 15:3 (2010), pp. 237-; DOI 10.1007/s10588-009-9059-1.
http://hdl.handle.net/10945/31216

Comput Math Organ Theory (2009) 15: 237–257
DOI 10.1007/s10588-009-9059-1

Efficient experimental design tools for exploring large simulation models

Regine Pei Tze Oh · Susan M. Sanchez · Thomas W. Lucas · Hong Wan · Mark E. Nissen

Published online: 6 November 2009

Abstract Simulation experiments are typically faster, cheaper and more flexible than physical experiments. They are especially useful for pilot studies of complicated systems where little prior knowledge of the system behavior exists. One key characteristic of simulation experiments is the large number of factors and interactions between factors that impact decision makers. Traditional simulation approaches offer little help in analyzing large numbers of factors and interactions, which makes interpretation and application of results very difficult and often incorrect. In this paper we implement and demonstrate efficient design of experiments techniques to analyze large, complex simulation models. Looking specifically within the domain of organizational performance, we illustrate how our approach can be used to analyze even immense results spaces, driven by myriad factors with sometimes unknown interactions, and pursue optimal settings for different performance measures. This allows analysts

R.P.T. Oh
DSO National Laboratories, Singapore, Singapore

S.M. Sanchez
Department of Operations Research, Naval Postgraduate School, Monterey, CA 93943, USA
e-mail: smsanche@nps.edu

T.W. Lucas
Department of Operations Research, Graduate School of Operational and Information Sciences, Monterey, CA 93943, USA
e-mail: twlucas@nps.edu

H. Wan (✉)
School of Industrial Engineering, Purdue University, 315 N. Grant Street, GRIS 256, West Lafayette, IN 47907-2023, USA
e-mail: hwan@purdue.edu

M.E. Nissen
Department of Information Sciences, Naval Postgraduate School, Monterey, CA 93943, USA
e-mail: MNissen@nps.edu

to rapidly identify the most important, results-influencing factors within simulation models, employ an experimental design to fully explore the simulation space efficiently, and enhance system design through simulation. This dramatically increases the breadth and depth of insights available through analysis of simulation data, reduces the time required to analyze simulation-driven studies, and extends the state of the art in computational and mathematical organization theory.

Keywords Simulation · Design of experiments · Screening · Hierarchy organizational model

1 Introduction

In the current economic climate, organizations must be agile, flexible, and responsive in order to survive. Innovative organizational structures are critical for organizational effectiveness, but identifying the most appropriate (e.g., non-hierarchical) organizational structures can be difficult. One way is via leveraging experts to identify a specific, alternate structure from organizational theory or practice; change one or more parts of the organization to mirror such structure; and then conduct some live experiments or exercises to assess the relative efficacy of the new structure. However, clearly, the cost and time required for this approach are huge, and depending upon the level of expertise driving the selection of the specific alternate structure chosen for organizational change, the new structure may perform no better—indeed, it may perform worse—than the original.

An alternative approach involves experimenting on computational models of organizations instead of operating organizations in practice. The basic idea is to save time and money by conducting experiments in the virtual as opposed to the physical domain and to use the power of simulation to examine the comparative performance of a large number of alternate organizational structures in a short amount of time. Key to such examination is the identification of the most important drivers of organizational performance, which can provide decision makers with rich insights into selecting the best organizational structure.

Researchers using simulation to explore organizational performance have coined the term "computational organizational theory" for this type of virtual experimentation (e.g., see Carley and Gasser 1999). Simulation models are often comprised of dozens or hundreds of model parameters, which can combine to generate immense results spaces. Such immense results spaces defy analysis by trial-and-error or full-factorial experimentation designs. New, more efficient designs and analytical approaches exist and continue to be improved, but they are largely unfamiliar to most researchers who conduct field, laboratory or computational experiments (Kleijnen et al. 2005). In particular, such researchers need efficient experimental designs and interfaces/infrastructure for automating experiments and the analysis of results data.

In this paper, we build upon substantial prior work along these lines to introduce and implement efficient design of experiments techniques to analyze large, complex simulation models. Looking specifically within the domain of organizational performance, we illustrate how our approach can be used to analyze even immense results spaces, driven by myriad factors with sometimes unknown interactions, and pursue optimal settings for different performance measures. This allows analysts to rapidly identify the most important, results-influencing factors within simulation models, employ an experimental design to fully explore the simulation space efficiently, and enhance system design through simulation. This dramatically increases the breadth and depth of insights available through analysis of simulation data, reduces the time required to analyze simulation-driven studies, and extends the state of the art in computational and mathematical organization theory.

The importance of such new capability to analyze computational experimentation results for simulations of complex organization is discussed in Gateau et al. (2007), who address the relative multidimensional performances of six theoretically distinct organizational forms. For example, as the authors state (p. 16),

> "*... if we can identify the model parameters that enable the Simple Structure to keep risk below that of the Machine Bureaucracy (e.g., formalization), that enable the Professional Bureaucracy to operate so quickly in predictable environments such as the Industrial Era (e.g., application experience), and that enable the Divisionalized Form to keep rework down in predictable environments (e.g., hierarchy)—that is, drawing the best from each organizational form—then we would establish the capability to design an organization that is tailored specifically to a particular environment. Further, if we can identify the model parameters that make each of the various organizational forms more or less effective in terms of responses to manipulations such as enhanced network architecture and increased professional competency, then we would establish the capability to design an organization that is tailored specifically to a particular manipulation. This represents the objective of articulating the organization design space: to facilitate organizational design specific to particular environments and managerial manipulation.*"

However, as we discuss in more detail in Sect. 2.2, the immense results spaces preclude easy identification of the most important model parameters and interactions using traditional experimental design techniques. In this paper, we describe new, state-of-the art experimental designs that allow analysts to overcome such difficulties with guaranteed correctness. For example, Controlled Sequential Bifurcation (CSB) is a new group screening method developed by Wan et al. (2006) that produces dramatic efficiency gains in large-scale computational experiments when the important factors are sparse. As another example, the related technique Fractional Factorial Controlled Sequential Bifurcation (FFCSB; see Sanchez et al. 2005, 2009; Oh 2007) extends the effectiveness of CSB in various, common modeling problems.

In the balance of this paper, we begin by discussing two experimentation tools used to build and test organization theory: (1) the POW-ER computational experimentation tool and (2) FFCSB. We then instantiate a computational model of a hierarchical organization, and we compare the opinions of subject matter experts with analysis via FFCSB. In this discussion, instantiation, and comparison, we illustrate the use and utility of the approach, and we show how it extends the state of the art. We close with a set of key conclusions and an agenda for continued research along these lines.

2 Experimentation tools for organization theory

In this section we draw heavily from Nissen (2007) to provide a brief overview of the tools used to represent computationally the Hierarchy organization. Much has been written elsewhere about the computational tools. The interested reader is directed to the corresponding citations and references below for additional information. We also include an appendix to supplement this detailed description, and to provide the interested reader with sufficient information to replicate our model. The goal is to enhance research reliability via such replication.

2.1 POW-ER computational experimentation tool

These complex computer simulations of organizational behavior are developed in POW-ER—Projects, Organizations and Work for Edge Research—a virtual environment for computational modeling of Command and Control (C2) organizations and processes. POW-ER builds upon collaborative research spanning nearly two decades to develop rich, theory-based models of organizational processes. Using an agent-based representation (Cohen 1992; Kunz et al. 1998), micro-level organizational behaviors have been researched and formalized to reflect well-accepted organization theory (Levitt et al. 1999). Extensive empirical validation projects (e.g., Christiansen 1993; Thomsen 1998) have demonstrated representational fidelity, and have shown how the qualitative and quantitative behaviors of POW-ER computational models correspond closely with a diversity of enterprise processes in practice.

The POW-ER modeling environment has been developed directly from Galbraith's (1977) information processing view of organizations. This information processing view has three key implications (Jin and Levitt 1996). The first is ontological: we model knowledge work through interactions of *tasks* to be performed, *actors* communicating with one another and performing tasks, and an *organization structure* that defines actors' roles and that constrains their behaviors. In essence this amounts to overlaying the task structure on the organization structure, and to developing computational agents with various capabilities to emulate the dynamic behaviors of organizational actors performing work. We model the organization structure as a network of reporting relations that can capture micro-behaviors such as managerial attention, span of control, and empowerment. We represent the task structure as a separate network of activities that can capture organizational attributes such as expected duration, complexity and required skills. Within the organization structure, we model further various *roles* (e.g., marketing analyst, design engineer, manager), which can capture organizational attributes such as skills possessed, level of experience and task familiarity. Within the task structure, we model further various sequencing constraints, interdependencies and quality/rework loops, which can capture considerable variety in terms of how knowledge work is organized and performed.

Also, each actor within the intertwined organization and task structures has a queue of information tasks to be performed (e.g., assigned work activities, messages from other actors, meetings to attend) and a queue of information outputs (e.g., completed work products, communications to other actors, requests for assistance). Each actor also processes such tasks according to how well the actor's skill set matches those required for a given activity, the relative priority of the task, the actor's work backlog (i.e., queue length), as well as how many interruptions divert the actor's attention from the task at hand. Collective task performance is constrained further by the number of individual actors assigned to each task, the magnitude of the task, and both scheduled (e.g., work breaks, ends of shifts, weekends and holidays) and unscheduled (e.g., awaiting managerial decisions, awaiting work or information inputs from others, performing rework) downtime.

The second implication is computational: both direct work (e.g., planning, design, operations) and indirect effort (e.g., rework, coordination, decision wait) are modeled in terms of *work volume*. This construct is used to represent a unit of work (e.g., associated with a task, a meeting, a communication) within the task structure. In addition to symbolic execution of POW-ER models (e.g., qualitatively assessing skill mismatches, task-concurrency difficulties, decentralization effects) through micro-behaviors derived from Organization Theory, the discrete-event simulation engine enables (virtual) process performance to be assessed quantitatively (e.g., projecting task duration, cost, rework, process quality).

The third implication is validational: the computational modeling environment has been validated extensively, over a period spanning almost two decades, by a team of more than 30 researchers (Levitt 2004). This validation process has involved three primary streams of effort: (1) internal validation against micro-social science research findings and against observed micro-behaviors in real-world organizations, (2) external validation against the predictions of macro-theory and against the observed macro-experience of real-world organizations, and (3) model cross-docking experiments against the predictions of other computational models with the same input data sets (Levitt et al. 2005; Orr and Nissen 2006, p. 8). As such, ours is one of the few, implemented, computational organization modeling environments that has been subjected to such a thorough, multi-method trajectory of validation.

Performance in POW-ER is measured through an array of the dependent variables: *completion time, cost, direct work, rework, coordination work, decision wait time, maximum backlog, and product risk*. In this study, we examine performance with respect to *completion time*, which is noted widely as one of the most important concerns of organizational managers, and which serves the purposes of this study well in terms of illustrating the use, utility and power of our approach. The examination of other performance measures represents a topic for future research along the lines of this investigation.

2.2 FFCSB: an efficient, adaptive approach for computational experimentation

Researchers employing computational models to represent organizations and simulate their behaviors typically used full factorial experimental designs to explore organizational performance. For instance, Nissen (2007) uses a 2 organizations × 2 scenarios design, while Gateau et al. (2007) employ a larger 6 organizations × 2 scenarios × 4 manipulations design. Rather than experimenting on individual factors, they use the six groups of factors listed above and change multiple factors within a group as one variation. Experimental results of organizational performance were analyzed over the entire organization's model changes and mission changes (Nissen

2007) or single block change (Gateau et al. 2007). Through analyzing the relative impact of each variation on individual organization performance, the researchers drew practical insights. For instance, Orr and Nissen inferred that: "professional competency improvements to the Hierarchy/Machine Bureaucracy can produce even more dramatic results in terms of agility as those associated with adopting the Edge organizational form. Hence, a change in professional competency can be substituted to a large degree for a change in organizational form. Unlike the substitution effects noted above for the network architecture manipulation, however, the converse does not hold for professional competency: changing organizational form does not compensate for a reversion to an efficiency-oriented organization and knowledge-flow approach" (2006, p. 16). Here "Hierarchy/Machine Bureaucracy" refers to the kind of rigid, hierarchical organization representative of most large and established firms (e.g., global corporations) and government agencies (e.g., militaries), and "Edge" refers to an alternate, flexible, non-hierarchical organization form noted as appropriate in today's dynamic environment (Alberts and Hayes 2003).

The approach of grouping factors into larger categories for experiments is one approach that researchers have taken to reduce the computational requirements; another is to restrict the number of factors to a very small number. What may not be obvious at first glance is how quickly it becomes computationally infeasible to conduct full factorial designs on individual factors—even for a single organization. For example, suppose that one of the organizational models has twenty factors that can be varied, the scenario has ten factors, and there are four manipulations. A full factorial design involving only low and high levels (i.e., no intermediate levels) for each factor for this single organization has $2^{20} \times 2^{10} \times 4$ runs per replication. Even if the computational model runs in one second, this requires 136 years of computer processing time! And, unless the response variances can reasonably be assumed to be constant, two or more replications are needed in order to estimate the effects with any statistical validity. Given that there are hundreds or thousands of factors in such complex organization models, it is easy to see why computational organizational researchers have limited their studies to a small number of factors or groups of factors. However, this also limits the types of insights that can be gleaned from a single simulation study.

Fortunately, FFCSB is an efficient, adaptive method that allows researchers to study large numbers of factors in a reasonable amount of time. FFCSB is an extension of the CSB group screening procedure (Wan et al. 2006) that adds a pre-screening stage to determine the sign of each factor's effect and pre-rank the effects. Rather than requiring the researcher to consolidate large sets of factors into a few composite categories (such as "industrial age" and "21st century") that can be examined using a full factorial design, FFCSB allows the researcher to efficiently assess the importance of all individual factor effects. This is key to providing detailed insights about how to make good use of investments to improve organizational performance. For example, certain aspects of the multi-dimensional category of professional competency may be very important, while others may have little impact. Organizations that invest in improving only a few dimensions of professional competency might focus on less-important aspects, hence fail to achieve significant improvements in performance; organizations that attempt to achieve in simultaneous improvement all aspects of professional competency might end up investing far more than necessary to

Initialization:

Specify the desired maximum Type I error α and the minimum power requirement γ.

Create two empty LIFO queues for groups, NEG and POS.

Phase 1:

Conduct a saturated or nearly-saturated fractional factorial experiment for the K individual factors and estimate $\hat{\beta}_1, \ldots, \hat{\beta}_K$.

Order the estimates so that $\hat{\beta}_{[1]} \leq \ldots \leq \hat{\beta}_{[z]} < 0 \leq \hat{\beta}_{[z+1]} \ldots \leq \hat{\beta}_{[K]}$.

Add the sorted factors $\{[1], \ldots, [z]\}$ to the NEG LIFO queue, and the sorted factors $\{[z+1], \ldots, [K]\}$ to the POS LIFO queue.

Phase 2:

For queue = POS and queue = NEG, do

While queue is not empty, do

Remove: Remove a group of size N from the queue.

Test:

Unimportant:

If the group is unimportant, then classify all factors in the group as unimportant.

Important (size=1):

If the group is important and of size 1, then classify the factor as important.

Important (size>1):

If the group is important and the size is greater than 1, then split the group into two subgroups, the first group of size $\lceil N/2 \rceil$, such that all factors in the first subgroup have smaller $[i]$'s (ordered indices) than those in the second subgroup.

Add each subgroup to the LIFO queue.

End Test

End While

End For

Fig. 1 Structure of FFCSB (adapted from Sanchez et al. 2009)

achieve better performance. An algorithmic description of FFCSB appears in Fig. 1. We provide a conceptual overview of FFCSB, followed by a more detailed technical discussion.

Conceptually, FFCSB begins by categorizing the individual factors into two groups based on rough estimates of their effects—positive or negative—on the response. This Phase 1 activity uses a single replication of an efficient fractional factorial design to sort the factors into two groups according to the sign of their estimated main effects. In Phase 2, one of two sequential group testing procedures is used to further delineate the factors: the procedure of Wan et al. (2006) called Controlled Bifurcation (CSB), or an extended version called CSB-X (Wan et al. 2009). In each step of CSB or CSB-X, the cumulative effect of a group of factors is tested for importance. If the group's effect is important, indicating that at least one factor in the group may have an important effect, then the group is split into two subgroups. The effects of these two subgroups are then tested in subsequent steps and each subgroup is either classified as unimportant or split into two subgroups for further testing. As the experiment proceeds, the groups become smaller until eventually all factors that have not been classified as unimportant are tested individually. The sequential property of the method makes it well suited for simulation experiments.

Numerical evaluations show that FFCSB is efficient and effective, especially when little or nothing is known about the factor effects (Sanchez et al. 2009). The performance of FFCSB arises from strong theoretical properties of CSB. CSB assumes a main-effects model with normal errors, and that all factor effects $\beta_i \geq 0$. CSB is the first procedure to control the probabilities of both types of misclassification error. Specifically, the error control of the sequential bifurcation procedure depends on the error control of each testing step. For a specific group of factors $\langle k_{l+1}, k_{l+2}, \ldots, k_m \rangle$, where k_i is the factor's index and β_{k_i} represents the effect coefficient of factor k_i, define a *qualified hypothesis test* as a test that guarantees:

$$\Pr\left\{ \text{Declare}\langle k_{l+1}, k_{l+2}, \ldots, k_m \rangle \text{ important} \,\middle|\, \sum_{i=l}^{m} \beta_i \leq \Delta_0 \right\} \leq \alpha; \quad \text{and}$$

$$\Pr\left\{ \text{Declare}\langle k_{l+1}, k_{l+2}, \ldots, k_m \rangle \text{ important} \,\middle|\, \sum_{i=l}^{m} \beta_i \geq \Delta_1 \right\} \geq \gamma.$$

The extended version CSB-X, also proposed by Wan et al. (2009), eliminates the effects of two-factor interactions; it is the first procedure to provide unbiased screening results for the main effects while providing these error guarantees. CSB is highly efficient for large-scale problems, particularly when a small proportion of the factors are truly important and these factors are clustered, since CSB can eliminate unimportant factors in groups. For example, Wan et al. (2006, 2009) demonstrate that the CSB and CSB-X procedures can save up to 75% of the runs required by one replication of a fractional factorial design, which in turn requires far fewer runs than a full factorial design. Another benefit of CSB and CSB-X is their validity when the error variances differ across different factor settings, as is commonly the case in simulation experiments. A single replication of fractional factorial design cannot give such a performance guarantee.

A drawback of CSB and CSB-X is that they require the signs of the factor effects to be known a priori. This is not realistic for many complex systems (Lucas et al. 2002). FFCSB eliminates the need for this assumption by relying on Phase 1 sampling—rather than expert opinion—to determine the signs of the effects. The inclusion of the fractional factorial design significantly reduces the probability that two critical effects cancel each other. In addition, by sorting the factors after the first phase, FFCSB significantly improves the efficiency of the second-phase CSB or CSB-X and can reduce the total computational effort (relative to CSB or CSB-X) by up to 64% (Sanchez et al. 2009) Although FFCSB is a heuristic procedure, extensive numerical evaluations show that FFCSB outperforms CSB and CSB-X in both efficiency and effectiveness, especially when little or nothing is known about the factor effects; FFCSB has been successfully applied to more than 86 000 experiments on test problems involving up to 1023 factors and a variety of different factor effect patterns and variance structures (Sanchez et al. 2005, 2009).

In short, FFCSB offers single factor resolution and allows researchers to probe questions such as: What are the most important factors, either organizational or mission, driving the measure of performance in an organizational model? Without group screening algorithms, it would have required an exorbitant amount of experimentation resources to conduct full factorial experiments to identify performance enhancement (or deterioration) due to single factors. FFCSB overcomes this limit by efficient division and experimentation of the entire factor space, and gradually limiting the scope of search for important factors. Through group screening of singular factors, FFCSB can shed light on significant individual factors within each structural or mission factor block that have the most impact on the outcome of interest.

3 Model description and significance

In this section we continue to draw from Nissen (2007) to describe the Hierarchy computational model. This model has been validated against an operational, military, Joint Task Force (JTF) organization (e.g., comprised of units from the Army, Navy and Air Force) in the field (see Nissen 2005). The organization is very familiar to the subject matter experts who contribute to our study.

3.1 Hierarchy organizational model

Figure 2 is a screen-capture of the Hierarchy model in the POW-ER environment. The figure illustrates the personnel hierarchy and mission structure in the Hierarchy model. Personnel are grouped and communicate over a 3-tier command chain, which emulates the Command, Coordination and Operations levels in a JTF Hierarchy (Nissen 2007). There are four tasks executed sequentially via two phases. Tasks are linked to each other and to project milestones. Tasks can flow completed work down the chain, or flow rework (additional work to rectify earlier mistakes). Personnel are linked to work on meetings and tasks. Operations level personnel act directly on tasks, while Command and Control level personnel act directly upon their specialized tasks while indirectly supporting operations tasks. Completion time is considered to be a key performance measure.

3.2 Factor exploration space

Table 1 lists the factors identified in the Hierarchy model for the FFCSB application. The desired Type I and Type II errors are both 0.05 (i.e., the desired power is 0.95). We divide the entire factor space into three subspaces for separate FFCSB exploration. Hence, three smaller and faster explorations are conducted instead of one big exploration. The division of the factor space follows the three manipulations of mission context factors: (1) mission and environmental context, (2) network architecture, and (3) professional competency. In addition, the three sets of structural factors: (1) organization structure, (2) communication structure, and (3) work structure are subsumed under these factor subspaces. This division of factor space is intended to mirror that in the literature as closely as possible, but is not exact. The factor ranges of exploration are derived from the default values of the Hierarchy model in the contrasting mission contexts of Industrial Age and 21st Century (see Appendix for details). In lieu of requesting subject matter experts to specify thresholds, we select these based on the range of effects observed in some preliminary experiments.

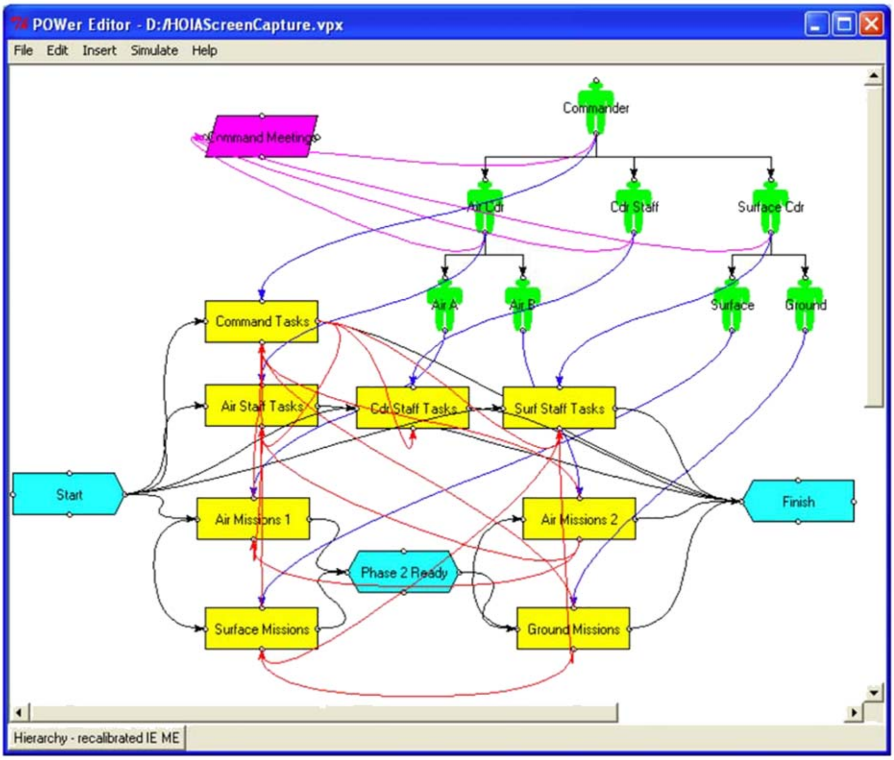

Fig. 2 Hierarchy organizational model in POW-ER (from Oh 2007)

FFCSB was also applied to the Hierarchy model with this entire factor space in one exploration. However, there were some unusually long simulation times and we cut off this single exploration after it ran for over a week without yielding final results.[1] The sequential nature of FFCSB meant that the experiments could not be parallelized. While we cut off this larger experiment, note that we still designed, conducted, and analyzed experiments involving a total of 115 individual factors over the course of a four-day workshop. For comparison purposes, full factorial experiments would have required 4.5 h per replication for the Network Architecture subspace, nearly 700 centuries for the Professional Competency subspace, and 366 billion centuries for the Mission & Environment subspace.

3.3 Expert opinion on significant factors

As we mention in the previous section, we did not make extensive use of subject matter experts in setting the factor ranges. One expert (who was also the model developer) identified four factors as important before the experiments began. Under Mission &

[1]Had we not already completed the exploration of the three subspaces, we could have let FFCSB run to completion, or stopped FFCSB and examined the partially-screened results.

Table 1 Factor space for exploration of hierarchy model

Mission & Environment (60 factors)	Network Architecture (14 factors)	Professional Competency (41 factors)
(Project) Function Exception Probability	(Project) Priority	(Project) Team Experience
(Project) Project Exception Probability	(Project) Length Of Work-day	(Personnel) Culture
(Task) Effort	(Project) Length Of Work-week	(Personnel) Role
(Task) Learning Days	(Project) Centralization	(Personnel) Application Experience
(Task) Priority	(Project) Matrix-strength	(Personnel) Cultural Experience
(Task) Requirement Complexity	(Project) Communication Probability	(Personnel) Skill Ratings
(Task) Solution Complexity	(Project) Noise Probability	
(Task) Uncertainty	(Project) Instance Exception Probability	
(Personnel) Full Time Equivalent	(Meeting) Priority	
(Personnel-Task) Allocation	(Meeting) Duration	
(Task-Task) Successor	(Personnel-Meeting) Allocation	
	(Task-Task) Rework Strength	

Environment, the two factors were (Personnel) Full Time Equivalent and (Task) Effort; under Professional Competency, the two factors were (Personnel) Application Experience and (Personnel) Skill Ratings.

We remark that it is useful to document the experts' expectations before running large-scale experiments—or at least ask about their expectations before revealing the results—to prevent opinions being set by the experimental results (Kleijnen et al. 2005). This helps focus later discussion by highlighting "surprises" in the results that may merit further investigation.

3.4 FFCSB findings on significant factors

Tables 2 and 3 summarize the FFCSB findings of important factors in the Hierarchy model that have the greatest impact on completion time. There were no factors classified as important in the Network Architecture factor subspace.

4 Discussion

Before we discuss the results, a few general comments are in order. First, all factors effects correspond to the impact on the *completion time* of changing that factor from its lowest to its highest value. Widening the range for a factor deemed unimportant in our experiment might make it show up as important, while narrowing the range

Table 2 Important factors in mission & environment factor subspace

Object	Attribute	Factor effect on completion time
Mission	Project Exception Probability	+
Surface Missions	Effort	+
Surface Missions	Solution Complexity	+
Ground Missions	Effort	+
Ground Missions	Requirement Complexity	+
Ground Missions	Solution Complexity	+

Table 3 Important factors in professional competency factor subspace

Object	Attribute	Factor effect on completion time
Mission	Team Experience	+
Air A (Personnel)	Skill Ratings	–
Ground (Personnel)	Skill Ratings	–

for a factor deemed important in our experiment might make it drop out. Similarly, an analyst using more stringent thresholds than ours to define what constitutes an important factor would tend to see fewer factors identified as important, while an analyst using a less stringent threshold would tend to see more. The goal of this study is not to provide a definitive assessment of how the Hierarchy model behaves in general, but rather to show that a large-scale screening experiment can provide a rich set of insights into the model's behavior. As well as a means of confirming or refuting specific hypotheses developed a priori, this can be used to focus discussion, generate additional hypotheses, or explore how robust the organizational performance is to variations in, say, the environment or the task.

4.1 Comparison of expert opinion and FFCSB results

As we mention in Sect. 3.3, we did not make extensive use of subject matter experts when developing the factor ranges. However, the model developer did identify several factors as important before the experiments began. When the results match expectations, this may provide the experts with assurance that the model as a whole is adequately capturing key relationships in the organizational structure and setting being modeled. When the results disagree with expert opinion, this focuses discussions on what to do next. If there is a problem with the model implementation, then software bugs need to be fixed. If "surprise" results occur because the model does not adequately reflect reality, then enhancements or changes may be necessary. However, if further inspection rules out these first two issues, then we have learned something new and expert opinion may need to adjust. Examples of all three of these outcomes appear in Kleijnen et al. (2005).

In the first factor subspace of Mission & Environment, the experts identified the factors of Full Time Equivalent (FTE) and Effort as important. FTE measures the equivalent of manpower resources available and Task Effort quantifies the time effort requirement of the task. Contrary to expert opinion, FFCSB does not classify any FTE factors as important over the factor range of exploration. Thus, FTE is not as important as the other factors in this subspace in impacting the completion time. This is an interesting finding, particular since FTEs were varied over a wide range (from one-half to twice the default number for each personnel category). The implications here are that the organizational performance is robust to the loss of FTE in any single personnel category, and that adding more manpower to any single group also has relatively little impact on completion time. Further experimentation could be used to confirm these results, or to estimate the net impact of simultaneous FTE changes in two or more teams.

In line with expert opinion, FFCSB classifies Effort factors as important, but of the eight possible missions, only those for Surface Missions and Ground Missions are flagged. Critical path analysis of the Hierarchy model explains why factors associated with only these two missions showed up consistently as important. The red bars in Fig. 3 depict the critical path of the project simulated in the Hierarchy model. Following the red bars, the Air Missions 1, Surface Missions and Ground Missions are on the critical path. Of these three missions, the Surface Missions and Ground Missions have minimum float, i.e., there is no allowance for shifting these missions in time. Hence, these two missions are crucial to achieving a rapid completion time. Besides the Task Effort factor, FFCSB also classified the Solution Complexity factors of the Surface and Ground Missions as important, as well as the Requirements Complexity of the Ground Missions. Thus, FFCSB has further quantified expert opinion by flagging only factors associated with missions on the critical path with specific characteristics.

In addition, FFCSB classifies the global factor of Project Exception Probability (PEP) as important. PEP is the probability that a subtask will fail and generate rework for failure dependent tasks. This factor is significant for the Hierarchy model that is characterized by sequential and interdependent tasks and hence, suffers a longer Project Duration in the event of increased PEP.

In the second factor subspace of Network Architecture, there are no factors classified as important. This finding is in agreement with the experts, who did not expect any important factors in this subspace. A set of (relatively computationally expensive) resolution V fractional factorial design (allowing the estimation of both main effects and two-way interactions) was used to verify the factor coefficients in this factor group. (Sanchez and Sanchez (2005) provide a simple method for generating resolution V fractional factorials for very large numbers of factors; code in the form of an executable jar file is available at the SEED Center for Data Farming web pages.[2]) The results confirmed that the factor coefficients were relatively small in magnitude and hence, practically insignificant. We remark that within the Department of Defense, Network Architecture is the area where the majority of the money and effort have gone in pursuit of developing net-centric or net-enabled forces. It is

[2]SEED (Simulation Experiments and Efficient Design) Center for Data Farming, Naval Postgraduate School. http://harvest.nps.edu/.

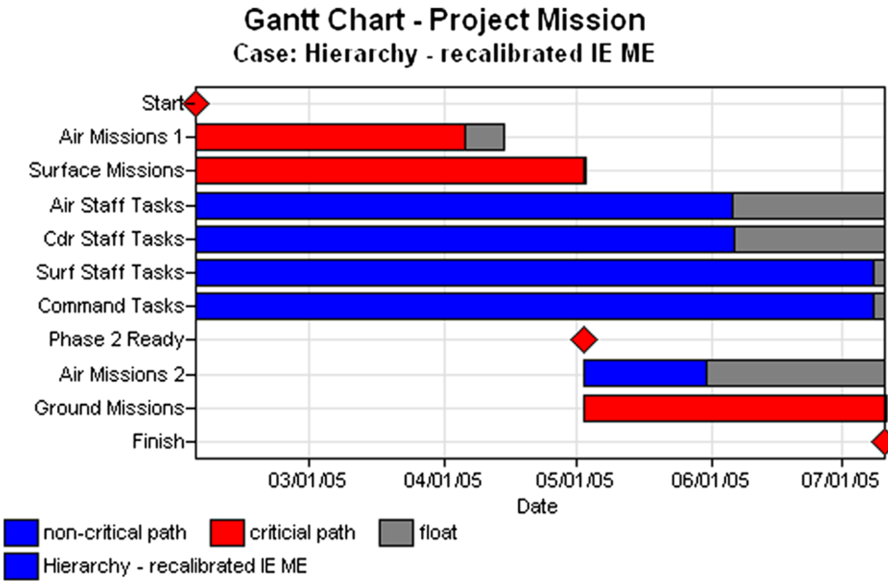

Fig. 3 (Color online) Critical path analysis of hierarchy model shows Air Missions 1, Surface Missions and Ground Missions on Critical Path

interesting that both our experts and our FFCSB results agree that Network Architecture is unimportant. These results suggest that, at least for Hierarchical command and control structures such as the one we study, a shift in focus to examine the mission environments and individual skill levels would be beneficial.

In the third factor subspace of Professional Competency, experts identified Skill Ratings and Application Experience factors as important. FFCSB classified the Skill Ratings of the Air A and Ground personnel as important, but not that of the Surface personnel. These three groups of personnel are operations personnel and directly responsible for the missions on the critical path. The contrast between the three missions is that the Surface Missions require a considerably longer effort of 21 months versus that of the Ground Missions (6.5 months) and Air Missions 1 (11 months). These findings suggest that Skill Levels may be more critical for missions that lie on the critical path and have relatively shorter Effort requirements. FFCSB did not classify Application Experience as important.

However, interestingly, FFCSB classified Team Experience as important and positively related to the measure of performance. Team Experience quantifies the degree of familiarity that team members have in working with one another as a team. In other words, this finding suggests that more team experience leads to longer Project Duration in the Hierarchy model. This result seems counter-intuitive. In fact, in two workshops subsequent workshops where participants were shown the factor categories in Table 1 and asked to identify factors they felt were important, Team Experience was chosen as a way to decrease completion time! Yet this seemingly counter-intuitive finding may have been observed in earlier research and experimentation. Ramsey and Levitt (2005) summarized high level findings from Horii et al. (2004) on the impact of cultural differences in project teams: "Japanese-style organizations were more effective, with either US or Japanese agents, at performing tasks with high interdependence when the team experience of members was low." The Hierarchy model studied in this application shares common characteristics of centralized authority, high formalization, and multiple hierarchies with the Japanese-style organization modeled in Horii et al. (2004, p. 3). In addition, these experiments had used Project Duration (i.e., completion time) and Quality Risk to quantify team performance, while this FFCSB application only used completion time. Hence, there is common ground to compare the similarity of both findings. Had the original intuition on Team Experience been applied with conventional screening algorithms, this factor could have distorted screening findings.

Lastly, there were two general observations of interest. First, there were more important factors associated with the Operations layer of the structure than the other layers. Recall that the Hierarchy model has a 3-tier command chain that models the Command, Coordination and Operations layers in a JTF. Second, there were more uncontrollable or difficult to control factors (e.g., Project Exception Probability, Task Requirement Complexity, Task Solution Complexity and Team Experience) than controllable or easy to control factors (e.g., Skill Ratings).

4.2 Choice of screening method

The important factor classification and observations are meant to provide direction for researchers in future work and optimize their experimentation budget on truly important factors. This first-case FFCSB application on a real-world simulation model has produced results that are coherent with critical path analysis and that agree with earlier research on similar models. Hence, it is an encouraging sign that FFCSB can serve as a complementary tool to better understand complex simulation models. Of course, these findings are preliminary and apply to a specific hierarchical command and control structure: care should be taken in drawing general conclusions.

FFCSB is not the only potential experimental design that can be applied to complex simulation models. Other experimental designs are also suitable for these types of applications, and further methodological work is currently underway. A variant called FFCSBX is useful for categorizing main effects even in the presence of two-way interactions (Sanchez et al. 2009). Another screening approach uses sequential fractional factorial designs with the same error control as CSB and is specifically fit for cases with non-sparse important effects (Shen and Wan 2005; Wan and Ankenman 2007). A hybrid approach discussed in Shen et al. (2009) allows the analyst to estimate factor effects (rather than simply classify factors as important or unimportant) at the completion of the experiment. Regardless of the screening procedure used, the analyst may wish to follow up with further experiments that examine those factors deemed important in more detail. For example, a new DOE-based algorithm is proposed in Chang et al. (2007), called Stochastic Trust Region Gradient-Free Method (STRONG), for solving large-scale simulation optimization problems since it is easy to automate and yet has provably reliable asymptotic performance. The screening experiment can significantly improve the efficiency of the STRONG procedure by reducing the dimension of the original problem. Another example involves those interested in multiple performance measures, since FFCSB and other screening methods are designed for single responses. Researchers might find it useful to run separate screening experiments for each performance measure, and determine whether or not the different performance measures are influenced by different sets of factors. Alternatively, they could use the results for the key performance measure to design secondary experiments (such as the familiar full factorials) involving a small number of factors, and then examine multiple responses. We reiterate that without an efficient screening method, those that only use a small-scale experiment must rely heavily on subject-matter expertise in narrowing down the factors, which may severely limit the insights into the system's behavior.

5 Concluding remarks

In this paper, we illustrate how an efficient experimental design approach can support current research in Computation Organization Theory. The FFCSB application produced many delightful surprises. Part of the important factor classification was in line with expert opinion and part of it ran contrary to expectations. There were new findings of important factors that were justified by critical path analysis and in agreement with earlier research and experimentation. Overall, this particular FFCSB application has confirmed expert opinion, flagged out new important factors and produced some interesting hypothesis, all for further exploration.

There are limitations to the FFCSB application to any model. FFCSB assumes a main effects model and interactions can distort the accuracy of factor classification. The nature of the response variance (homogeneous or heterogeneous) and its magnitude are unknown. Both model characteristics can have bearings on the FFCSB findings and accuracy guarantees. Particular to the Hierarchy model, the observations of this FFCSB exploration are unique to the factor space organization and ranges of exploration. Hence, the findings are not conclusive of the Hierarchy model. The important factor classification and observations are meant to provide direction for researchers in future work and optimize their experimentation budget on truly important factors. This first-case FFCSB application on a real-world simulation model has produced results that are coherent with critical path analysis and that agree with earlier research on similar models. Hence, it is an encouraging sign that FFCSB can serve as a complementary tool to better understand complex simulation models.

Continued exploration of the Hierarchy model with different factor space organization and factor ranges would form a good sensitivity analysis study of the FFCSB application on the model. Exploring an Edge organization model would form an interesting study in itself, and allow for meaningful contrasts between the competing organizational forms.

The benefits of being able to easily perform a screening experiment on a complex organizational model cannot be overstated. In the absence of this capability, an analyst must either limit themselves to a small number of factors to investigate, or make changes to a large number of factors simultaneously to come up with a small number of organizational forms, settings, or task types to investigate. We remark that the POW-ER model and its predecessor have been successfully used in practice for over a decade. Although this model that has been "validated" by a history of successful applications in the field, it is nonetheless difficult for experts to fully grasp the complex interplay of the complete set of potential factors. This is particularly important in Command and Control research, as we seek—not to model existing organizations and organizational structures—but to define new ones that will be effective for our military transformation.

Screening experiments also offer opportunities to validate a model for a particular use. For example, if results contradict expert opinion and, after further discussion or field experiments, the model results are shown to be inaccurate, the model should be modified. Controversial results from a screening experiment may, in fact, help identify alternatives that merit testing in the field. In the long run, this cycle of model-test-model will lead to models that provide better representations of reality, as well as a better understanding of the model's behavior, strengths, and limitations.

In summary, there are efficient experimental designs and screening approaches that are easy to implement, require fewer assumptions than conventional experimental design methods, and yet can provide analysts with better insights when the experiments are complete. These can substantially reduce the computational requirement for military leadership to identify optimal factor combinations, and lead to a much broader and deeper understanding of the system. This will facilitate the decision making process dramatically.

Acknowledgements We thank Marc Ramsey for guidance on running POW-ER in a batch mode, and Steve Upton for assistance in linking to POW-ER from the FFCSB program. This work was supported in part by grants from the *Center for Edge Power*[3] at the Naval Postgraduate School, by the *Modeling and Simulation Coordination Office* for the Joint Test and Evaluation Methodology project, and by *Netcentric Systems Test/Science and Technology* program.

Appendix

Table 4 lists the factors identified in the Hierarchy model for FFCSB application and their range of exploration. Note that some factors are varied individually, such as separate factors for every personnel role and every task type.

[3]Center for Edge Power, Naval Postgraduate School. http://www.nps.edu/Academics/Centers/CEP/index.html.

Table 4 Factors & ranges in hierarchy model

Object	Factor	Organizational Structure		FFCSB Exploration	
		Industrial Age Hierarchy	21st Century Hierarchy	Low Value	High Value
Project	Priority	Medium	Medium	Low	High
	Work-day	480	480	360	600
	Work-week	2400	2400	1440	3600
	Team Experience	Low	Low	Low	Medium
	Centralization	High	High	Medium	High
	Formalization	High	High	Medium	High
	Matrix Strength	Low	Low	Low	Medium
	Communication Probability	0.1	0.1	0.05	0.2
	Noise Probability	0.3	0.3	0.01	0.6
	Functional Exception Probability	0.1	0.2	0.05	0.4
	Project Exception Probability	0.1	0.2	0.05	0.4
	Instance Exception Probability	0	0	0.01	0.4
Meeting	Priority	High	High	Medium	High
	Duration	2 hours	2 hours	0.5 hours	4 hours
Personnel	Culture	Generic	Generic	American	Japanese
	Role	(Various)	(Same)	PM	ST
	Application Experience	Medium	Low	Low	Medium
	Cultural Experience	Medium	Medium	Low	High
	Full Time Equivalent	(Various)	(Same)	0.5* Default	2* Default
	Skill Ratings	Medium	Medium	Low	High
Task	Effort	(Various)	(Same)	0.5* Default	2* Default
	Learning Days	0	0	0	90
	Priority	Medium	Medium	Low	High
	Requirement Complexity	Medium	High	Medium	High

Table 4 (*Continued*)

Object	Factor	Organizational Structure		FFCSB Exploration	
		Industrial Age Hierarchy	21st Century Hierarchy	Low Value	High Value
	Solution Complexity	Medium	High	Medium	High
	Uncertainty	Medium	High	Medium	High
Meeting Assignment		0.1–1.0		0.1	1.0
Task Assignment	Allocation	0.9–1.0	1.0	0.7	1.0
Successor	Time Lag	0	0	0.0 pct-complete	0.5 pct-complete
Rework	Strength	(Various) 0.15, 0.3, 1.0	0.1	0.15	0.3

References

Alberts DS, Hayes RE (2003) Power to the edge: command… control… in the information age. Command and control research program (CCRP) publication series. Department of Defense, Washington

Carley K, Gasser L (1999) Computational organizational theory. In: Weiss G (ed) Distributed artificial intelligence. MIT Press, Cambridge, Chap. 7

Chang K-H, Hong JL, Wan H (2007) Stochastic trust region gradient-free method (STRONG)—a new response-surface-based algorithm in simulation optimization. In: Proceeding of the 2007 Winter Simulation Conference, Dec 9–12, 2007, pp 346–354. Available online at http://www.informs-sim.org/wsc07papers/041.pdf. Accessed January 10, 2008

Christiansen TR (1993) Modeling efficiency and effectiveness of coordination in engineering design teams. PhD thesis, Department of Civil and Environmental Engineering, Stanford University

Cohen GP (1992) The virtual design team: an object-oriented model of information sharing in project teams. PhD thesis, Department of Civil and Environmental Engineering, Stanford University

Galbraith JR (1977) Organization design. Addison-Wesley, Reading

Gateau JB, Leweling TA, Looney JP, Nissen ME (2007) Hypothesis testing of edge organizations: modeling the C2 organization design space. In: Proceedings international command & control research & technology symposium, Newport (June 2007). Available at http://www.nps.navy.mil/GSOIS/cep/docs/2007/ICCRTS07_Gateau.pdf. Accessed January 10, 2008

Horii T, Jin Y, Levitt RE (2004) Modeling and analyzing cultural influences on team performance through virtual experiments. Available at http://crgp.stanford.edu/publications/conference_papers/TamakiCMOT.pdf. Accessed October 23, 2007

Jin Y, Levitt RE (1996) The virtual design team: a computational model of project organizations. Comput Math Organ Theory 2(3):171–195

Kleijnen JPC, Sanchez SM, Lucas TW, Cioppa TM (2005) A user's guide to the brave new world of designing simulation experiments. INFORMS J Comput 17(3):263–289

Kunz JC, Levitt RE, Jin Y (1998) The virtual design team: a computational simulation model of project organizations. Commun Assoc Comput Mach 41(11):84–92

Levitt RE (2004) Computational modeling of organizations comes of age. J Comput Math Organ Theory 10(2):127–145

Levitt RE, Orr RJ, Nissen ME (2005) Validating the virtual design team (VDT) computational modeling environment. The collaboratory for research on global projects. Working Paper Series #25, pp 1–15

(September 2005). Available at: http://crgp.stanford.edu/publications/working_papers/WP25.pdf. Accessed January 10, 2008
Levitt RE, Thomsen J, Christiansen TR, Kunz JC, Jin Y, Nass C (1999) Simulating project work processes and organizations: toward a micro-contingency theory of organizational design. Manage Sci 45(11):1479–1495
Lucas TW, Sanchez SM, Brown L, Vinyard W (2002) Better designs for high-dimensional explorations of simulations. In: Horne G, Johnson S (eds) Maneuver warfare science 2002, USMC Project Albert, Quantico, pp 17–46. Available via http://www.projectalbert.org/files/MWS2002On-line.pdf. Accessed July 1, 2005
Nissen ME (2005) Hypothesis testing of edge organizations: specifying computational C2 models for experimentation. In: Proceedings international command & control research symposium, McLean, VA, June 2005. Available at http://www.nps.navy.mil/GSOIS/cep/docs/2005/Nissen64-ICCRTS2005.pdf. Last accessed January 10, 2008
Nissen ME (2007) Computational experimentation on new organizational forms: exploring behavior and performance of edge organizations. Comput Math Organ Theory 13(3):203–240
Oh RPT (2007) Fractional factorial controlled sequential bifurcation: efficient screening through divide and discard. MS thesis, Operations Research Department, Naval Postgraduate School, Monterey, CA
Orr RJ, Nissen ME (2006) Computational experimentation on C2 models. In: Proceedings international command and control research and technology symposium. Cambridge, UK (September 2006). Available at http://www.nps.navy.mil/GSOIS/cep/docs/2006/Orr-Nissen057-ICCRTS2006.pdf. Accessed January 10, 2008
Ramsey M, Levitt RE (2005) A computational framework for experimentation with edge organizations. In: Proceedings international command & control research & technology symposium, McLean, VA (June 2005). Available at http://www.nps.navy.mil/GSOIS/cep/docs/2005/Ramsey-Levitt-ICCRTS2005.pdf. Accessed January 10, 2008
Sanchez SM, Sanchez PJ (2005) Very large fractional factorial and central composite designs. ACM Trans Model Comput Simul 15(4):362–377
Sanchez SM, Wan H, Lucas TW (2005) A two-phase screening procedure for simulation experiments. In: Proceedings of the 2005 winter simulation conference, pp 223–230. Available at http://www.informs-sim.org/wsc05papers/023.pdf. Accessed March 23, 2007
Sanchez SM, Wan H, Lucas TW (2009) A two-phase screening procedure for simulation experiments. *ACM Trans Model Comput Simul* 19(2): Article 7
Shen H, Wan H (2005) Controlled sequential factorial design for simulation factor screening. In: Proceedings of the 2005 winter simulation conference, pp 467–474. Available at http://www.informs-cs.org/wsc05papers/052.pdf. Accessed October 23, 2007
Shen H, Wan H, Sanchez SM (2009) The hybrid approach for simulation factor screening. Nav Res Logist (to appear)
Thomsen J (1998) The virtual team alliance (VTA): modeling the effects of goal incongruency in semi-routine, fast-paced project organizations. PhD thesis, Department of Civil and Environmental Engineering, Stanford University
Wan H, Ankenman BE (2007) Two-stage controlled fractional factorial screening for simulation experiments. J Qual Technol 39(2):126–139
Wan H, Ankenman BE, Nelson BL (2006) Controlled sequential bifurcation: a new factor-screening method for discrete-event simulation. Oper Res 54(4):743–755
Wan H, Ankenman BE, Nelson BL (2009) Improving the efficiency and efficacy of CSB for simulation factor screening. *INFORMS J Comput* (to appear)

Regine Pei Tze Oh was born in Singapore in 1978. She clinched a Defence Technology Training Award in 1997 to attend University of Toronto, Canada. In 2001, Regine graduated with a Bachelors of Applied Science (Electrical Engineering) and received the W.S. Wilson Medal for achievement of the highest standing in the final year. Upon graduation from University of Toronto, Regine joined DSO National Laboratories as a research engineer. In 2006, Regine embarked on the joint Masters program offered by Temasek Defence Science Institute. She has completed her Masters of Defense Science & Technology at the National University of Singapore and her Masters of Science in Operations Research at the Naval Postgraduate School. After graduation from Naval Postgraduate School, Regine returned to family and friends in Singapore and is contributing towards the defence community in Singapore.

Susan M. Sanchez is a Professor in Operations Research at the Naval Postgraduate School, and Co-Director of the Simulation Experiments & Efficient Design (SEED) Center for Data Farming. She also holds a joint appointment in the Graduate School of Business & Public Policy. She has a B.S. in Industrial & Operations Engineering from the University of Michigan, and a Ph.D. in Operations Research from Cornell. She has been active in various service capacities within the simulation community over many years, and is currently on the WSC Board of Directors. Her web page is http://faculty.nps.edu/smsanche/ and her e-mail is ssanchez@nps.edu.

Thomas W. Lucas is an Associate Professor of Operations Research at the Naval Postgraduate School (NPS) in Monterey California. He has been at NPS for eleven years. Previously, he worked for six years as a statistician and project leader at RAND and for eleven years as a systems engineer at Hughes Aircraft Company. A detailed CV can be found at: http://www.nps.navy.mil/orfacpag/resumePages/lucas.htm. His primary research interests are simulation, warfare analysis, design of experiments, and robust Bayesian statistics. In addition to his regular faculty duties, Professor Lucas is a co-director of NPS's SEED Center for Data Farming-see http://harvest.nps.edu. SEED is an acronym for Simulation Experiments and Efficient Design. The Center's mission is to advance the collaborative development and use of simulation experiments and efficient designs to provide decision makers with timely insights on complex systems and operations.

Hong Wan is an Assistant Professor in the School of Industrial Engineering at Purdue University. She is currently visiting the Department of Operations Research at Naval Postgraduate School. Her research interests include design and analysis of simulation experiments, simulation optimization, applied statistics, quality management, and healthcare engineering. She has taught a variety of courses and is a member of INFORMS and ASA. She is currently serving as the associate editor of ACM Transactions on Modeling and Computer Simulation. Her email address is hwan@purdue.edu and her web page is http://web.ics.purdue.edu/~hwan.

Mark E. Nissen is Chair Professor of Command & Control, and Professor of Information Science and Management, at the Naval Postgraduate School. His research focuses on dynamic knowledge and organization for competitive advantage. He views work, technology and organization as an integrated design problem. His 2006 book, entitled *Harnessing Knowledge Dynamics: Principled Organizational Knowing & Learning*, reflects this view and elaborates Knowledge Flows Theory and applications. Mark's extensive publications span information systems, project management, organization studies, knowledge management, counterterrorism and related fields. In 2000 he received the top Naval Postgraduate School faculty research award. In 2001 he received a prestigious Young Investigator Grant Award. In 2002–2003 he was Visiting Professor at Stanford. In 2004 he established the Center for Edge Power. Before earning his PhD from the University of Southern California, he acquired over a dozen years' management experience in the aerospace and electronics industries.