Reducing uncertainty in Systems Engineering through Defence Experimentation

Shane D. Arnott
PhD Student
School of Information Technology & Electrical Engineering
The University of Queensland

Professor Peter A. Lindsay
Professor of Systems Engineering
School of Information Technology & Electrical Engineering
The University of Queensland

ABSTRACT

Defence Experimentation (DE) is becoming adopted in Defence as a means to improve requirements elicitation, especially in capability development. However there is currently no established method for predicting the effort required in DE. This paper makes some first steps to address this, by carefully surveying the benefits of DE to the Systems Engineering process, the different needs of DE and the nature of the simulation methods and tools involved. It then surveys the literature on forecasting and prediction, with particular attention to methods from software engineering, where the problem has been tackled with some success. A parametric modelling approach to DE effort prediction is proposed.

INTRODUCTION

The reduction of uncertainty in Systems Engineering projects is an eternal goal. Systems Engineering is employed in the face of complexity, where an interdisciplinary approach is required to achieve a successful system. A continuing area of weakness in the Systems Engineering process is evident in the early stages, centring on the step from elicitation of customer needs into a defined problem statement. Getting this step right is particularly important in Defence, where the cost of getting it wrong can be enormous.

Defence Experimentation has been established as a means to combat this weakness by providing greater interaction with the customer towards developing an improved understanding of the problem, particularly in capability development and deployment of innovative solutions. However the discipline is relatively young, and as yet there is no well established method for estimating the cost and schedule of experimentation that is likely to be required, which in turn makes it difficult to integrate this activity into the overall project schedule.

Systems Engineering: where are we going wrong?

The Standish Group has collected statistics on information technology projects over the last 15 years, and their findings paint a bleak picture of project success rates (Standish-Group 2009). For example, in 2009, only 32% of the projects surveyed met the criteria for success (i.e. completed on time, on budget, and with all the features originally specified). Of the remainder, 44% were challenged (i.e. late, over

To appear: Proceedings Improving Systems and Software Engineering Conference (ISSEC), Brisbane, Australia, August 2010.

budget or lacking in features) and the final 24% failed completely (i.e. cancelled or delivered and never used).

While the aforementioned failure rates are often repeated information, Standish went further to identify success factors elicited from respondents to the project surveys. Of the top ten major success or failure factors, the majority are related to the early stages of the Systems Engineering process (i.e. user involvement, clear statement of requirements, minimized scope and realistic expectations).

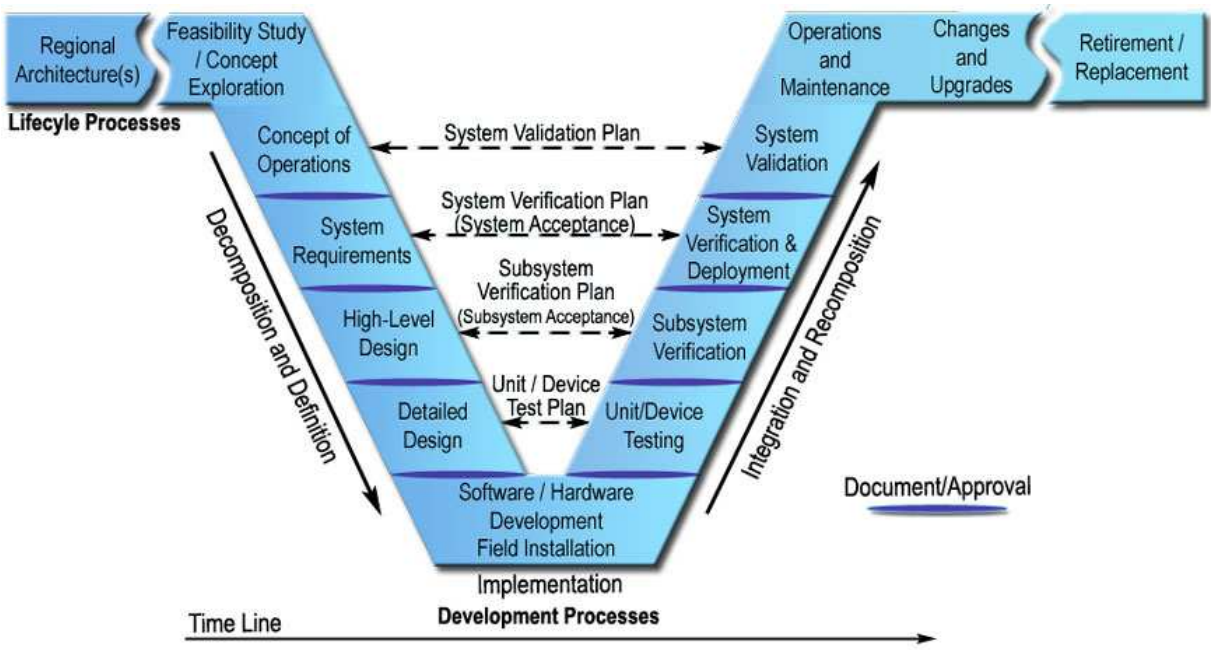

Figure 1. Systems Engineering V process (Federal-Highway-Administration 2007)

The Systems Engineering V process is depicted in this way to show the symmetry and dependency of respective development processes towards completion - meaning the left hand side needs to be performed in such a way to enable the right hand side to produce a successful result. As cited by the Standish report, often the left hand side is not being adequately performed for the right hand processes to function effectively.

Defence Experimentation is focused on the tasks involved on the left hand side (i.e. decomposition and definition) specifically on the initial phases of concept exploration, feasibility assessment, concept of operations and systems requirements definition.

This paper aims to define a framework for Defence Experimentation, and its utility in improving the customer-needs-to-problem-statement process, in order to better estimate the costs (including schedule costs) involved. After surveying the different approaches to Defence Experimentation, and the nature of the simulation methods and tools employed, the focus turns to forecasting and prediction methods for experimentation campaigns. Such methods are important in order better integrate Defence Experimentation into the Systems Engineering process, so as to establish improved problem definition into overall Systems Engineering schedules.

DEFENCE EXPERIMENTATION

Departments of Defence in many developed nations have embraced the activity of

Defence Experimentation in order to improve the understanding of concepts and establishment new or improved defence force capabilities. The targeted activities are often grouped under the term of "capability development".

> **Defence Experimentation** is "the application of the experimental method to the solution of complex defence capability development problems, potentially across the full spectrum of conflict types, such as warfighting, peace-enforcement, humanitarian relief and peace-keeping" (TTCP 2005).

The history of Defence Experimentation is short, with the concept being born out of the Command and Control Research Program (CCRP) Information Age Transformation series on better decision making for defence, specifically two "early" seminal publications in Experimentation: the Code of Best Practise for Experimentation (COBP-E) (Alberts and Hayes 2002) and Campaigns of Experimentation (Alberts and Hayes 2005). These publications note the high importance of experimentation to the future of Defence. The COBP-E preface remarks that "experimentation is the lynch pin in the [US] Department of Defense's strategy for transformation. Without a properly focused, well-balanced, rigorously designed, and expertly conducted program of experimentation the Department of Defense will not be able to take full advantage of the opportunities that Information Age concepts and technologies offer".

The Technical Cooperation Program (TTCP) Joint Systems Analysis (JSA) Action Group 12 published a document authored by leading experts in the area of experimentation from America, Britain, Canada and Australia titled the Guide for Understanding and Implementing Defence Experimentation (GUIDEx) (TTCP 2005). This more recent document takes the basis laid by (Alberts and Hayes 2002; Alberts and Hayes 2005; Kass 2006) and develops it further, by focusing on providing guiding principles to practitioners and those more closely related with the experimentation process. As the GUIDEx points out, this type of decision support has become even more critical in recent times: "The development of allied forces has always been a difficult and complex process. However the need for force development to respond to asymmetric and unpredictable threats, the demands of coalition operations, the perceived need for information supremacy, combined with evolving transformational technologies and concepts, has caused this task to become even more difficult over the past few years. Experimentation offers a unique means to support the development and transformation of allied forces by advancing our knowledge of the complex networked systems and capabilities likely to be fielded in the near future." (TTCP 2005)

More specifically the direct benefits of experimentation are "[the ability to] deliver timely answers with a measured level of confidence, thereby contributing to sound risk management of programs and their components. It thoroughly supports defence problem solving from concepts through capability development to operations" (TTCP 2005).

Employment of Defence Experimentation

The Code of Best Practise for Experimentation (COBP-E) defines three major uses of experimentation:

Discovery – to determine the efficacy of something previously untried;

Hypothesis testing – to examine the validity of a hypothesis; and

Demonstration – to examine and demonstrate a known truth.

Discovery experimentation involves introducing novel systems, concepts, organisational structures, technologies or other elements in a setting where their use can be observed and catalogued. In the Defence context the objective is to find out how the innovation is employed and whether it appears to have military utility as well as the limiting conditions – situations where the benefits may not be available. In a scientific sense these are "hypothesis generation" efforts that will be typically employed early in the development cycle.

A disadvantage of Discovery experimentation is that it will not ordinarily provide enough information or evidence to reach a conclusion that is valid (correct understandings of the cause and effect or temporal relationships that are hypothesised) or reliable (can be recreated in another experimentation setting). The advantages are that most new concepts, ideas, and technologies will benefit by using Discovery experimentation as a way of weeding out ideas that simply do not work, forcing the community to ask rigorous questions (facilitation of experts) about the benefits being sought and the dynamics involved in implementing the idea, or specifying the limiting conditions for the innovation.

Hypothesis testing is the classic experimentation used by scholars to advance knowledge by seeking to falsify specific hypotheses (specifically if-then statements) or discover their limiting conditions. In order to conduct hypothesis testing experiments, the experimenters create a situation in which one or more factors of interest (dependent variables) can be observed systematically under conditions that vary the values of factors thought to cause change (independent variables) in the factors of interest, while other potentially relevant factors (control variables) are held constant, either empirically or through statistical manipulation.

A disadvantage of this type of experimentation is that since the number of independent, dependent and control variables relevant in the military arena are very large, considerable thought and care is often needed to conduct valid hypothesis tests. Moreover no single experiment is likely to do more than improve knowledge marginally and help clarify new issues. Hence "sets of hypothesis testing" are often needed in order to gain useful knowledge.

Demonstration experiments are not intended to generate new knowledge, but rather to display existing knowledge to people unfamiliar with it [education]. In such demonstrations, all the technologies are well established and the settings (scenario, participants etcetera) are orchestrated to show that these technologies can be employed efficiently and effectively under the specified conditions.

Campaigns of Defence Experimentation

In addition to single experiments, the concept of a sequence of related experiments is described as an "experimentation campaign" (or simply "campaign" as it will be referred to from here on in). The reason for campaigns, as declared by the COBP-E, is that "military operations are too complex and the process of change is too expensive for [any country] to rely on a single experiment to prove that a particular innovation should be adopted". The ability to design and conduct individual experiments constitutes a necessary but not sufficient core capability to conceive, design, and conduct successful campaigns of experimentation (Alberts and Hayes 2005). Thus multiple experiments aimed at the same problem are required.

The GUIDEx notes a driving reason for conducting campaigns is that "using a variety of techniques ensures that weaknesses in one technique can be mitigated by others. Where the results (inferences) correlate between activities, it increases confidence and where they diverge, it provides guidance for further investigation. It is only when all activities are brought together in a coherent manner and the insights synthesised, that the overall problem under investigation is advanced as a whole." The technique is realised in the model-test-model experimental process, the foundation of the Australian Army Battlelab process (Bowley and Lovaszy 1999; ADF 2000; Bowley, Castles et al. 2003).

An illustration of how an idealised campaign of experiments may be sequenced along a "campaign vector" is presented below. The three dimensions (x, y, z) representing complexity of the experiment, the level of "use" (as discussed in the previous Employment of DE section) and method of experimentation (discussed in the next section of DE Methods) respectively. The notion is to start with simple discovery style war-games and to head towards complex demonstrations within live military exercises in order to explore and understand a new capability.

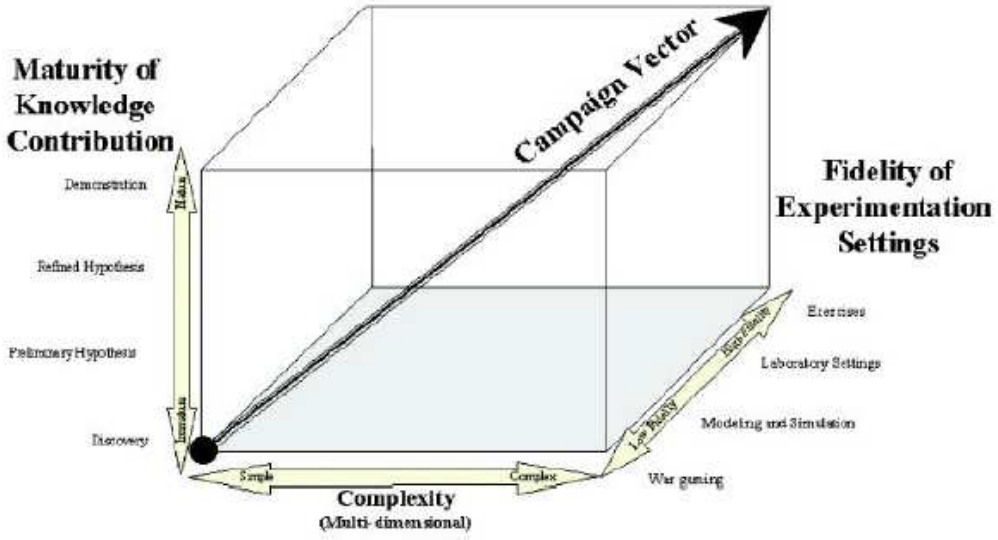

Figure 2. Defence Experimentation Campaign vector. (Alberts and Hayes 2002)

DEFENCE EXPERIMENTATION METHODS

The GUIDEx classifies Defence Experimentation activities through the use of technology (primarily simulation) into one of the following four "methods":

Constructive simulation, Analytic wargames, Human-in the-loop simulation, and Live simulation (field) experiments (TTCP 2005; Kass 2006).

Constructive simulations are those in which no human intervention occurs in the play after designers choose the initial parameters and then start and finish the simulation. Constructive simulations are a mainstay of military analytical agencies. They allow repeated replay of the same battle under identical conditions, while systematically varying parameters such as the insertion of a new weapon or sensor characteristic, or the employment of a different resource or tactic, or the encounter of a different threat. Experiments using constructive simulations with multiple runs are ideal to detect change and to isolate its cause. Because modelling complex events requires many assumptions, including those of variable human behaviour, critics often question the applicability of constructive simulation results to operational situations.

Analytic wargames typically employ command and staff officers to plan and execute a military operation. At certain decision points the Blue players give their course of action to a neutral, White cell, which then allows the Red players to plan a counter move, and so on. The White cell adjudicates each move, using a simulation to help determine the outcome. A typical analytic wargame might involve fighting the same campaign twice, using different capabilities each time. The strength of such wargames for experimentation resides in the ability to detect any change in the outcome, given major differences in the strategies used. Additionally, to the extent that operational scenarios are used and actual military units are players, analytic wargames may reflect real-world possibilities. A major limitation is the inability to isolate the true cause of change because of the myriad of differences found when attempting to play two different campaigns against a similar reactive threat.

Human-in-the-loop simulations represent a broad category of real-time simulations with which humans can interact. In a human-in-the-loop defence experiment, military subjects receive real-time inputs from the simulation; make real-time decisions and direct simulated forces or platforms against simulated threat forces. The use of actual military operators and staff allows this type of experiment to reflect warfighting decision-making better than experiments using purely constructive simulation. However, when humans make decisions variability increases and changes are more difficult to detect and consequently more difficult to attribute to the cause.

Live simulation is conducted in the actual environment, with actual military units, equipment and operational prototypes. Usually only weapon effects are simulated. As such, the results of experiments in these environments, often referred to as field experiments, are highly applicable to real situations. Good field experiments, like good military exercises, are the closest thing to real military operations. A dominant consideration however, is the difficulty in isolating the true cause of any detected change, since field experiments include much of the uncertainty, variability, and challenges of actual operations; but they are seldom replicated due to costs.

Choosing the "right" method and related support technology is crucial. The recognised measure of simulation merit for a particular problem is of its *fitness for purpose*. The definition for this term is as follows:

> **Fitness for purpose** is achieved by providing the capabilities, correctness, accuracy and usability needed for the intended use or current application (DMSO 2001)

The COBP-E notes the "important consideration in [choosing] models of complex systems is the identification of what are the essential properties of the system. These will be situation [or problem] dependant, and should be carefully aligned with the goals of the experiment" (Alberts and Hayes 2002).

In order to assist in method/tool selection the GUIDEx provides four requirements in designing a valid experiment. They are:

Requirement 1, Make use of the new or alternate application of a capability
Requirement 2, Detect a change in the effect
Requirement 3, Isolate the reason for the change
Requirement 4, Relate the results to actual operations

Each GUIDEx method of experiment provides different coverage of these four validity requirements and no single method provides perfect validity. Thus re-enforcing the need of combination approaches as theorised by the campaigns of experimentation to capitalise on the strengths on varied techniques.

For example the ability for precision and control increases the ability to detect change and to isolate its cause (requirements 2 and 3), although, in turn, it decreases the ability to apply the results to imprecise, real world situations (requirement 4). The mapping of GUIDEX methods versus the validity requirements is depicted below:

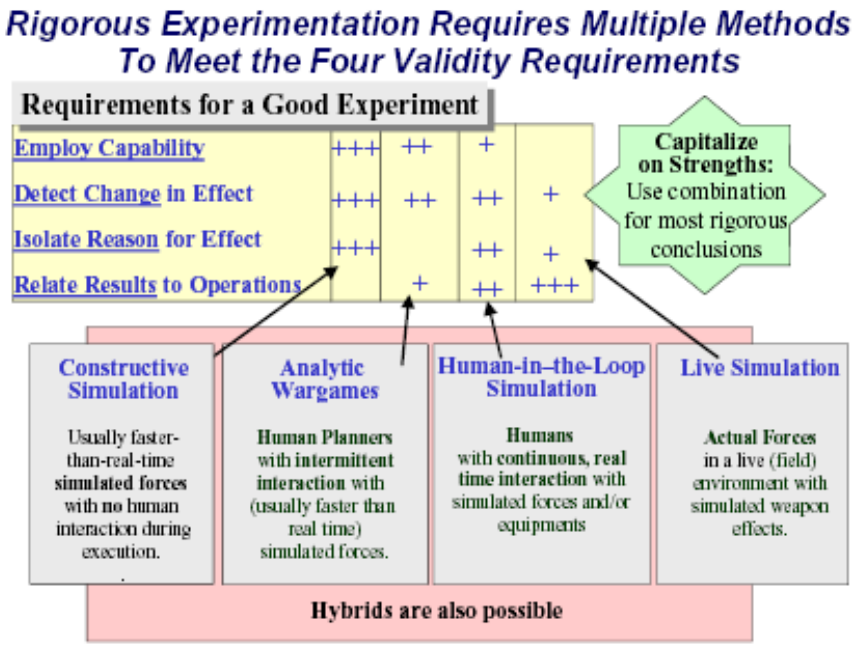

Figure 3. GUIDEx four validity requirements versus methods (TTCP 2005)

Putting it all together, if we follow the GUIDEx process, using the four validity requirements in our design to choose the right method (or campaign of methods), the types of outcomes that are advertised to result include:

- A richly crafted capability development package that clearly defines the innovation (i.e. the new or alternate application of a capability) and the elements necessary for its success;
- A set of research results that form a coherent whole and specify how the innovation should be implemented, the cause and effect relationships at work, the conditions necessary for success, and the types of military benefits that can be anticipated;
- A community of interest that includes researchers, operators, and decision makers who understand the innovation and are in a position to assess its value; and
- A significant reduction in the risks associated with adopting the innovation.

FORECASTING AND PREDICTION

The motivation for conducting Defence Experimentation is clear. That is, the gravity of getting the Systems Engineering definition phase "wrong" is large enough to spend additional effort and resources on Defence Experimentation activities to gain a greater understanding of the problem in order to better execute Systems Engineering projects inline with customers needs and concepts of operations.

Much is written about the theory (Alberts and Hayes 2002; Alberts and Hayes 2005) and practise (TTCP 2005; Kass 2006) of Defence Experimentation. But very little is found on the "cost of performing Defence Experimentation". That is, there is a distinct lack of formal planning aids to answer key questions of any Project Manager or Chief Engineer such as: "How much time needs to be allocated to better understand my customer's needs?" and "What type of resources will I need and how long will it take?" That is, how do I plan for success?

These are reasonable and important questions that are faced by any Management and planning team, whose function is to balance resources available against the constraints of schedule, budget and (acceptable) risk. Given the supporting nature of Defence Experimentation the importance of being able to execute such activities ahead of key decision milestones is a critical path issue and, to be useful, must be well understood. Making future predictions based on limited knowledge is known as forecasting (Armstrong 2001), that can take many forms as depicted below based on the type of information available at the time of planning.

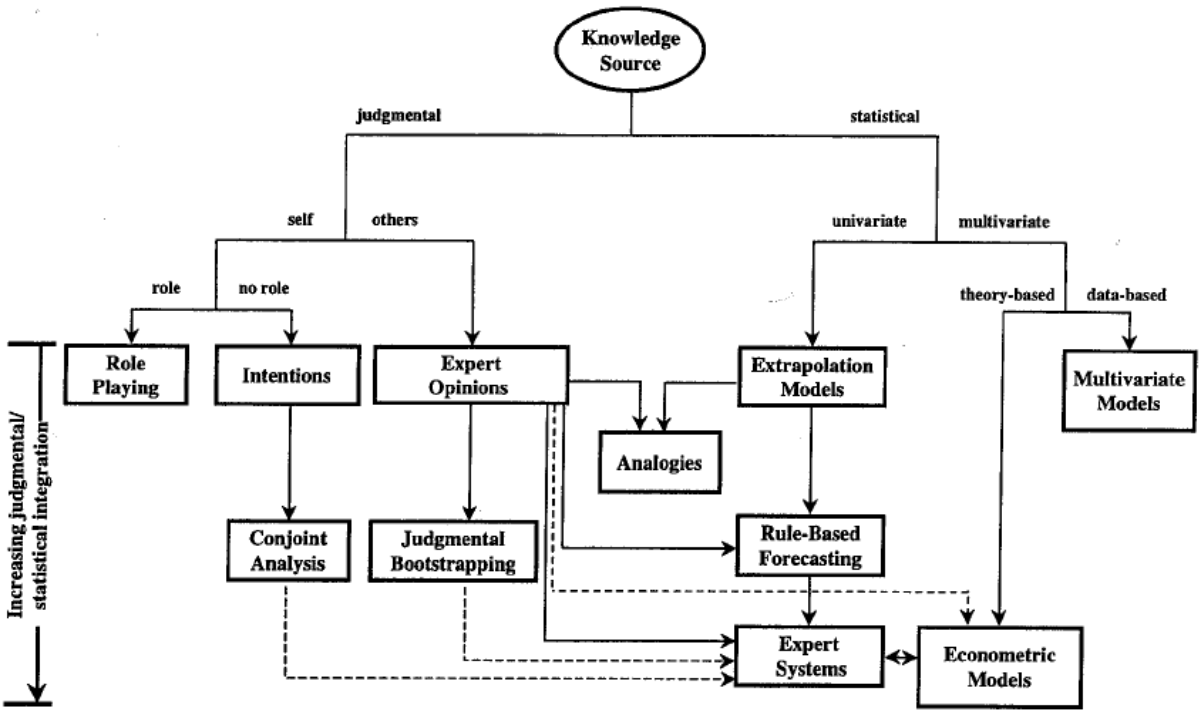

Figure 4. Forecasting Methodology Tree: Characteristics of forecasting methods and their relationships (Armstrong 2001)

As we see, there are two primary types of forecasting, one relying on judgement (or expertise) and the other determined on a statistical basis. A description of each technique is discussed in (Armstrong 2001). Each method has its own strengths and weakness with general criticisms across the major types being that judgemental methods require experts with relevant experience to be available and statistical methods require a valid dataset to the developmental context (i.e. type of projects, organisation etc).

Software Engineering Estimation

With the intent to develop a method to better forecast Defence Experimentation activities, we researched methods in similar disciplines that had a similar need for planning estimation understanding in order to increase effective execution. The field of Software Engineering had the most significant set of papers, with techniques in this discipline being adapted for similar fields such as Systems Engineering. A notable survey of software development effort estimation methods (Jørgensen 2004) provides mapping to the methodology tree above:

Approach	Forecasting type	Examples
Analogy-based estimation	Judgemental → Structured analogies	ANGEL (Shepperd, Schofield et al. 1996)
Bottom up estimation	Judgemental → Decomposition	Work breakdown structure
Group estimation	Judgemental → Expert forecasting	Delphi (Boehm 1984)
Parametric models	Statistical → Causal methods	SLIM, COCOMO, SEER-SEM (Putnam 1978; Boehm 1984; Galorath and Evans 2006)
Size Estimation models	Statistical → Extrapolation models	Function point analysis Story point (Jørgensen 2004)

Table 1. A summary of popular software development estimation approaches (Jørgensen 2004)

The authors' review of the software engineering literature seeking the "best" approach for estimation within Software Engineering was inconclusive. The dominant discussion taking place between parametric and judgemental methods involves multiple papers each claiming one method was more useful than the other. With this disagreement regarding which is the "best" method, another angle was pursued by the author to see what techniques were the most researched with the assumption that this was a good initial method to follow in order to give Defence Experimentation some sort of foundation for effort forecasting.

The table below presents such a study.

Estimation	-1989	1990-1999	2000-2004	Total
Regression	21 (51%)	76 (47%)	51 (51%)	148 (49%)
Analogy	1 (2%)	15 (9%)	15 (15%)	31 (10%)
Expert Judgement	3 (7%)	22 (13%)	21 (21%)	46 (15%)
Work breakdown	3 (7%)	5 (3%)	4 (4%)	12 (4%)
Function Point	7 (17%)	47 (29%)	14 (14%)	68 (22%)
Classification and regression trees	0 (0%)	5 (3%)	9 (9%)	14 (5%)
Simulation (role play)	2 (5%)	4 (2%)	4 (4%)	10 (3%)
Neural network	0 (0%)	11 (7%)	11 (11%)	22 (7%)
Theory	20 (49%)	14 (9%)	5 (5%)	39 (13%)
Bayesian	0 (0%)	3 (2%)	2 (2%)	5 (2%)
Combination of estimates	0 (0%)	3 (2%)	2 (2%)	5 (2%)
Other	2 (5%)	7 (4%)	16 (16%)	25 (8%)

Table 2. Analysis of software estimation papers written per approach (Jørgensen and Shepperd 2007) – note one paper may discuss more than one approach, hence the overall percentage imbalance

By far the most dominant approach have been regression based techniques, with approximately half of all research focused on this technique.

What does seem to represent somewhat of a consensus is the need to combine multiple methods in order to address the weaknesses of one approach with the strengths of another. In particular combining a statistical with a judgemental forecasting method such as parametric and analogy based estimators seems to yield the best result (Boehm 1984). With this in mind and given the distinct lack of (published) estimation methods in Defence Experimentation, it is important to start somewhere in order to work towards more scientific planning methods. Given the large body of research in the regression approaches this seems like the sensible place to start.

Additionally, despite the criticisms of the statistical forecasting approaches using parametric models not giving quality estimates all of the time (Shepperd, Schofield et al. 1996), it can be argued these models played an important role in maturing the field of Software Engineering. Such models provide a foundation for recording effort data on real projects and provide a basis to enable the argument of what is the best method. Parametric modelling methods have contributed significantly to the identification of cost drivers and have had significant impact on effort, cost and schedule estimation.

FUTURE WORK

Based on the assessment that a parametric estimation method would be an important contribution for greater utilisation of Defence Experimentation within Systems Engineering, an effort to collect case studies is underway to gain a greater understanding of the cost drivers associated with such activities. To date three major organisations have provided support to this on-going research, granting access to over 50 case studies for parametric profiling. These organisations reside in Australia and the United Kingdom, with customers of their services being the Australian Defence Force and UK Ministry of Defence respectively. Of note, studies from US Government sources could not be included due to export control constraints on release of the data to the (open) academic domain.

These organisations are characterised as:

1. A large multinational organisation that conducts Defence Experimentation for its own purposes in aid of market research, customer engagement, product development and as a contract to the Government conduct concepts development.
2. A Government Scientific organisation that conducts Defence Experimentation for the purposes of project requirements definition and tactics development
3. A Government sponsored industry consortium that provides Defence Experimentation as a service to the Government on a range of issues from concept development to project requirements definition and tactics development.

This case study database is providing the foundation for the development of a parametric estimation model through regression analysis of this real world historical data. The intent of this model is to provide an effective planning aid to project planners and managers estimating the effort associated with Defence Experimentation in support of wider processes such as Systems Engineering. This model is known as the “Defence Experimentation COst MOdel” (DECOMO) and will be published in due course.

SUMMARY AND CONCLUSIONS

The importance of Defence Experimentation is becoming increasingly recognised, and standards of best practice are appearing. This paper presented a summary of the literature on the matter and outlined the different types of experimental campaigns typically employed. However, it is apparent that no methodology has yet been established for objectively estimating the effort involved. We outlined the issues involved and surveyed methods used for estimating effort in large software engineering projects. We discussed which aspects of these methods are relevant to estimation of Defence Experimentation effort, and which would be appropriate for adaptation and extension to provide an accurate objective forecasting approach to Defence Experimentation. In conclusion the direction of future research will centre on the creation of a regression based parametric Defence Experimentation effort prediction model, with the purpose of better integration of Defence Experimentation into the overall Systems Engineering plan. This is intended to provide improved problem definition phase towards increased project success.

REFERENCES

ADF (2000). The Army Experimental Framework, Australian Army, Australian Defence Force.

Alberts, D. S. and R. E. Hayes (2002). Code Of Best Practice for Experimentation (COBP-E). Washington DC, Command and Control Research Program.

Alberts, D. S. and R. E. Hayes (2005). Campaigns of Experimentation: Pathways to Innovation and Transformation. Washington DC, Command and Control Research Program.

Armstrong, J. S. (2001). Principles of Forecasting, Kluwer Publishing.

Boehm, B. W. (1984). "Software engineering economics." IEEE Transactions on Software Engineering **10**(1): 4-21.

Bowley, D., T. Castles, et al. (2003). "Constructing a SUITE of Analytical Tools: A Case Study of Military Experimentation." ASOR Bulletin **22**(4): 2-10.

Bowley, D. and S. L. Lovaszy (1999). Use of Combat Simulations and Wargames in Analytical Studies. SimTect.

DMSO (2001). VV&A Recommended Practices Guide - Glossary, Defence Modelling and Simulation Office, US DoD.

Federal-Highway-Administration (2007). Systems Engineering for Intelligent Transport Systems. US-DoT, US Department of Transportation.

Galorath, D. and M. Evans (2006). Software Sizing, Estimation, and Risk Management. Boca Raton, FL, Auerbach Publications.

Jørgensen, M. (2004). "A review of studies on expert estimation of software development effort." Journal of Systems and Software **70**(1-2): 37-60.

Jørgensen, M. and M. Shepperd (2007). "A Systematic Review of Software Development Cost Estimation Studies." IEEE Transactions on Software Engineering **33**(1): 33-52.

Kass, R. (2006). The logic of warfighting experiments. Washington DC, Command and Control Research Program.

Putnam, L. (1978). "A General Empirical Solution to the Macro Software Sizing and Estimation Problem." Transactions on Software Engineering **4**(4): 345-361.

Shepperd, M., C. Schofield, et al. (1996). Effort estimation using analogy (ANGEL). 18th International Conference on Software Engineering,, ICSE.

Standish-Group (2009). Chaos Report: Information Technology Project Success Research. Boston, Standish Group.

TTCP (2005). Guide for Understanding and Implementing Defense Experimentation (GUIDEx). Washington DC, Technical Cooperation Program (TTCP), Joint Systems Analysis (JSA) Group, Methods and Approaches for Warfighting Experimentation Action Group 12 (AG-12).