18th ICCRTS:
"C2 in Underdeveloped, Degraded and Denied Operational Environments"

Paper ID: 48

Meta-Analysis of Multiple Simulation-Based Experiments

Topics: Topic 5: Experimentation, Metrics, and Analysis
Topic 1: Concepts, Theory, and Policy
Topic 6: Modeling and Simulation

Authors:

François Bernier, Ph. D.
Defence R&D Canada – Valcartier
2459 blvd. Pie-XI North
Quebec, QC, G3J 1X5
Canada
francois.bernier@drdc-rddc.gc.ca
+1 418 844 4000

David S. Alberts, Ph. D.
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA, 22311
USA
dalberts@ida.org
+1 703 845 2411

Paul Pearce
Defence Science and Technology
Laboratory
Fort Halstead,
Sevenoaks, Kent, TN14 7BP
United Kingdom
pvpearce@dstl.gov.uk
+44 01959 89 2087

Point of contact:
François Bernier, Ph.D.
francois.bernier@drdc-rddc.gc.ca
+1 418 844 4000

Abstract

Computer simulations facilitate the study of real world phenomena by providing safe, controlled, flexible, and repeatable experimental environments at costs far lower than other options. Both the validity of the simulation model and the scope of its validity determine the degree to which the findings can be generalized and applied to real world situations and problems. In practice, no single model captures all of the important aspects of a phenomena of interest, nor is applicable over a wide set of missions and circumstances. Thus, effectively utilizing a variety of models in a prospective meta-analysis (a set of common hypotheses and controllable variables and comparable metrics) offers the opportunity to improve validity and extend the findings to a broader range of real world situations. NATO SAS-085, a research group exploring C2 Agility and Requisite Maturity, conceived and developed an international meta-analysis approach for studying various aspects related to C2 Agility from multiple simulation-based experiments. This paper presents the methodology they employed which was inspired by the prospective meta-analysis domain. The challenges that arose from differences among these experiments, differences in the ways C2 Approaches were instantiated, and differences in the measures of success and in the conditions they considered are discussed.

1 Introduction

Computational science has provided new and powerful tools in the last few decades (Humphreys, 2004) that enable us to conduct more cost effective, less destructive, better controlled and more repeatable experiments. Massive computation and big data analytics (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011) are but two recent examples. Although conducting experiments and more recently simulation-based experiments is commonplace, combining data from more than one experiment (not simply looking at the findings) is a more recent development. Mega-analysis, pooled analysis, and meta-analysis (Bravata & Olkin, 2001) are three methods for combining data and/or results from various experiments (Curran & Hussong, 2009). The ability to increase the sample size and variety of data generated within an experiment has many advantages, including increasing statistical power, reducing exposition to local biases, and for the meta-analysis introducing better control for between-study variations.

Meta-analysis is a method that combines the results of multiple experiments with the objective of identifying patterns, similarities and discrepancies among the results. The meta-analysis approach is usually retrospective, i.e. it is based on already published studies, and uses high level findings such as the effect sizes as opposed to the data itself, since such data is usually not available. Contrary to real-world experiments (e.g. subject-based) where conducting an experiment can be costly, simulation-based experiments devote their resources to developing the model, not in collecting or analyzing these data. Repeating a simulation-based experiment for testing different hypotheses or reusing the simulation model (experimental setup) in another experiment are often much efficient approaches than doing the same with a real-world experiment. Consequently, simulation-based experiments offer the possibility of designing experiments based on a meta-analysis approach before those experiments are conducted. Such an approach is called prospective meta-analysis.

The meta-analysis, which employs data generated from multiple simulation models used in various experiments, is an adaptation of prospective meta-analyses conducted in human and life sciences (Ghersi, Berlin, & Askie, 2011) to the domain of computer simulation. Simulation based experiments offer the ability to explicitly control the environment and manipulate independent variables in such a way that it becomes possible to repeat an experiment under a large range of different conditions at minimal cost. However, a particular instantiation of simulation model is limited in a number of ways, e.g. it may have a limited number of dependent and independent variables to draw on, it may have a reduced scope or it may be slow running limiting its utility for exploring the problem space. Using a set of simulation models instead of just one allows the analyst to consider more possibilities. The advantages of using a prospective meta-analysis are the same as those of retrospective meta-analyses but, in addition, because it is designed before the experiments are conducted, it produces data that are more likely to be comparable rather than drawing on the data available from a retrospective meta-analysis, i.e. combining the findings of multiple past experiments. In addition, a prospective meta-analysis offers the opportunity to exploit the potential of the raw data which is not possible when combining high level results. In a prospective meta-analysis, since hypotheses are identified in advance, it becomes possible to generate data that are relevant and more complete for the selected set of hypotheses to be tested than it would be otherwise.

A few domains have successfully applied meta-analyses to computer simulations. Multi-model climate is a good example with multi-model ensemble (Tebaldi & Knutti, 2007) and Coupled Model Intercomparison Project (CMIP) (Meehl, Boer, Covey, Latif, & Stouffer, 2000). However, there are two issues related to the concept of a meta-analysis, first, it is not specifically defined for computer-based experiments, second, its use is uncommon among the simulation community. As a result, there are no guidelines that describe how to conduct a meta-analysis involving a number of simulation-based experiments. This paper aims at applying this practice, found in another domain of research, to simulation-based experiments.

A meta-analyses of multiple experiments must adhere to the same design process employed for a singular simulation-based experiment (Barton, 2004) but with additional considerations. Those considerations refer to the hypotheses formulation, the selection of independent and dependent variables, the elaboration of the experimental design and the statistical models, and the analysis of results. This paper devotes a section to each aspect of the experimental design and conduct. The benefits of meta-analyses are then illustrated using some of SAS-085's experimental results and analysis findings.

2 Meta-Analyses of Simulation-Based Experiments

2.1 Why Meta-Analyses

Combining many experiments into a single integrated one provides important advantages compared to an extensive review of already published studies, or even a meta-analysis of existing studies.

Undertaking a meta-analysis of multiple experiments offers the following benefits:

- **Generalization**: a meta-analysis potentially increases the generalizability of the results by ensuring the uniformity in the hypotheses and in the variables is accounted for, while promoting exploration of a diversity of contexts with a range of different models. Not only are results of a meta-analysis applicable to the study space that includes all of the circumstances that are considered in the set of model runs conducted, but they are also applicable to all of the in-between contexts not explicitly tested (potentially a virtually infinite number of (sub)contexts that could have been created or chosen for this purpose).
- **Cross-Platform Results**: a meta-analysis offers better control for between-experiment variations by explicitly considering variations in the fixed and random effects within the modelling due to the different instantiations of context and common independent variables. Thus, differences in results that would appear in various independent experiments are subtracted/removed, leading to more uniform, general, and meaningful results.
- **Increased Statistical Tests**: the meta-analysis increases the power of statistical tests that rely on the sample size[1] by combining data from many experiments. For instance, when the sample size is small, the differences observed cannot be established as not arising from random variations and thus the test will not be sufficiently discriminating.
- **Reduced Individual and Local Biases**: a meta-analysis reduces the influence of local biases. For instance, individual experimenters can choose inappropriate measures or unconsciously choose those that support their theories or the model that they employ may be biased towards favoring certain outcomes. Another potential source of error is that individual models or experiments could be open to criticism in some way. For example, they may contain errors in the implementation of the simulation model or make oversimplified assumptions, make errors in data capture or mistakes during data manipulation which could bias the experiment and produce lower quality results. In a meta-analysis, these "random" unintentional errors are expected to cancel each other out, either partially or entirely, and thus produce less biased and higher quality results. A side effect of combining error is to increase variability and confound main effects with between-experiment variability, therefore a proper statistical model was chosen for dealing with this variability. A statistical model for dealing with this variability is presented in Section 2.4.
- **Promote Synergies, Interactions and Discussions Among Researchers**: A more subtle benefit of a meta-analysis is to favor interactions and discussions as well as the setting of common goals among multiple researchers. Designing the meta-analysis and conducting experiments in collaboration is more likely to create fruitful interactions and better orient future research. In addition, the meta-analysis approach fosters highly critical thinking, helps challenge assumptions, and supports the

[1] Statistical power is proportional to the square root of the sample size.

generation of insights leading to proposals for alternative assumptions. Designing and conducting a meta-analysis provide a formal and rigorous way to revisit theory and hypotheses that contrasts with the usual white board exercises and talking where these systems are less restrictive and do not force their users to be consistent. In an example experienced by the authors, even if researchers agreed on a hypothesis, they still had their own "internal" interpretation and the underlying concepts it conveys. Talking was not enough to express one's internal interpretation and probably lead to more confusion. A very effective approach to agree on the interpretation of a hypothesis is to proceed with the next step, i.e. with the experimental design including the method of analysis. This approach provides a more formal way of expressing ideas and helps to clarify often elusive concepts. Of course researchers will always disagree at some level on what must be done, but at least they agree on what they disagree about. The outcome of this approach is not only a better designed set of experiments and associated meta-analysis but also a better shared understanding of the concepts under study.

2.2 Selecting Simulation Models and Developing Hypotheses for a Meta-Analysis

A simulation-based experiment exploits a single simulation model that is usually verified and validated for a given domain of applicability and a limited set of experimental conditions (Sargent, 1994). When a single experiment instantiates a model in order to test some hypotheses, verifications are made to ensure that the conditions of validity are respected. Consequently, any arbitrary simulation model cannot be used for an experiment and then included in a meta-analysis just for the sake of improving statistical power. In addition, the independent and dependent variables captured by individual simulation models can vary considerably. Finally, models represent somewhat different realities and perspectives and are suitable to test different (but hopefully related) hypotheses. These differences among models make it challenging to combine them in a meta-analysis. However, there are ways to meet these challenges and thus, take advantage of the opportunities that meta-analyses provide.

The solution strategy for selecting simulation models must be pragmatic in that there is a need to scope the analysis to enable it to be accomplished within available resource constraints. There are a number of ways in which the selection could be undertaken. A waterfall (or top-down) process is one possible approach where a top-down design process begins with establishing the objectives of the meta-analysis and identifying the specific hypotheses that will be explored. This approach provides a sound basis for selecting among existing simulation models[2] whose validity has been established. This approach is rarely workable in practice because it assumes either little restriction on the conditions of validity of the simulation models, a large number of simulation models, or simulation models with conditions of validity compatible with the aims of the meta-analysis. The latest reason is probably the most frequently encountered.

Utilizing an iterative process is a more flexible option. During a first iteration, general objectives and candidate hypotheses are defined, then suitable experiments are identified and available simulation models are assessed to determine their validity for supporting the objectives of the meta-analysis. These assessments include the ability of the experimental platforms to manipulate variables of interest and to generate measures of interest. Once this assessment is completed the objectives and hypotheses are revisited and a further refinement is undertaken, including the addition of more hypotheses, based on the improved understanding as to the capabilities of the available simulation models. Certain simulations models will lack data to test some hypotheses of even be incompatible for selected hypotheses. That is, not all of the models will contribute to all of the objectives of the meta-analysis. Capitalizing on the strengths of the available experimental platforms, while minimising the effect of any weaknesses, is the most challenging aspect of the design and conduct of the meta-analysis.

[2] A simulation model is used to conduct an experiment and then both terms can often be used interchangeably, except when the same simulation model is used in more than one experiment under different conditions/configurations.

The meta-analysis design process is not immune to the selection bias called file drawer problem, a phenomenon well known in realm of meta-analysis (Egger & Smith, 1998; Sterne, Egger, & Smith, 2001). It was found that published studies have a positive bias because studies with negative outcomes are less likely to be published. In addition, a side effect of publications is to publicize in some circles the simulation models on which the studies were conducted. Other valid models would then be harder to find and less likely to be included in a meta-analysis. Another issue is that simulation models are not selected purely randomly in any of these two approaches, a condition required by most statistical tests. However, it is not the selection of the model themselves that matters but the totality of their treatment of the variables of interest. In any case, it is difficult to definitively establish the conditions required for most statistical tests. However, common sense would lead one to the conclusion that a meta-analysis is more likely to generate data that is more representative of a larger population than using a single model.

In many fields, the number of existing simulation models that are applicable to a specific meta-analysis is very limited and thus there is a tendency to work with the limitations that exist rather than rejecting models that are less than well suited. As stated earlier, one needs to carefully consider testing some hypotheses with a subset of models or a subset of the data generated. If time and resources permit, modifying some simulation models, is an option, but needless to say this must been done carefully by individuals that really understand these models so that changes do not invalidate the model. Changes that simply increase the granularity of the information captured are a good option.

2.3 Defining Common Independent and Dependent Variables

To facilitate the merging of data from each experiment it is necessary to undertake the important task of predefining and documenting the independent and dependent variables with the aim of establishing a clear audit trail and ensuring a common understanding. The first step consists of deciding which dependent variables are needed to test the hypotheses and which independent variables are appropriate for determining their effect on the dependent variables. The viability of measuring any particular dependent variable of interest depends on the ability of the simulation models to instantiate the independent values and measure the dependent variables of interest. In some cases, simulation models can measure independent and dependent variables using different scales and an understanding has to be established to determine the degree of correspondence across these difference scales. Normalization across the scales can help to mitigate differences in the way variables are measured across the simulation models. The modelling of effects, described in the experimental design section, provides an additional and even more efficient way to manage these differences. As for selecting appropriate variables and the range(s) of values that they can take on, there are a few ways to make the task easier and add rigour in the process.

One approach resides in identifying theories and definitions that reflect the concepts underlying a variable. For instance, there is an important corpus of literature about how situational awareness should be measured. And this corpus explains how each measure relates to each other. In another example, the NEC C2 Maturity Model, or N2C2M2 (Alberts, Huber, & Moffat, 2010), describes what a C2 Approach should be and then elaborates on a few criteria that describe each level a C2 Approach can take. This theory can be used to compare the levels of C2 Approaches implemented by different simulation models and retain those that comply with theory.

A second approach to simplifying the task is to consider if variability is preferable to uniformity. In the previous example, the most common approach to testing hypotheses related to one or more C2 Approaches requires having essentially the same instantiation of each C2 Approach in each model. However, it may be better to foster variety in other situations. For instance, testing the agility of an organization is usually accomplished by measuring how well it performs against a wide range of challenges, the sum of which constitutes an endeavor space. The endeavor space is usually simulation model specific because it depends on the situation being simulated (e.g. a degraded network for network centric warfare-related simulation). The amount of variability and variety that is inherent in a meta-analysis can be far greater than if one utilizes one model since a set of simulation models will cover a wider variety of challenges for evaluating agility. The resulting design is a variable nested within the simulation model (endeavor space within simulation model).

In another example, the distribution of information (DoI), which refers to the extent to which the information needed to accomplish required tasks is available to each participant, is a central concept that must be measured. Any good metric that aims to capture the essence of the DoI concept would certainly have to incorporate many aspects related to DoI, such as timeliness, completeness, accuracy, etc. Each experiment has to consider how to metricate the dimensions of the C2 Approach Space with each capturing a different aspect of these concepts which becomes an advantage. Although this is not as good as having every experiment measuring every aspect of these concepts it is better than relying on a single narrow measure or no measure at all.

2.4 Modeling Effects

It is important to establish an explicit statistical model (not to be confused with a simulation model) for the meta-analysis that provides the foundation for a meta-analysis. The purpose of a statistical model is to establish relationships between and among the variables of interest, the validity of which is important for the hypotheses under test. Experimental results not only serve to sustain/disprove hypotheses but also help to improve the statistical model by estimating values for parameters. When some of these independent variables are probabilistic, a statistical test must be employed. The family of statistical models (e.g. linear regression) and tests (e.g. student t) available is vast. The choice of which statistical models and tests to use depends on the number and types of dependent and independent variables, the type of distribution of values observed for dependent variables, and the relationship between and among variables (linear, quadratic). Some statistical models are more general, like the generalized linear mixed model and regression model. To complicate the task of selecting the proper technique, they are designated differently according to the domain of application. Determining the equivalent of a participant in “within participant experiment” or of repeated in “repeated measures experiment” for a simulation-based experiment requires some level of knowledge about statistics. This paper presents two important and generic statistical models that were selected for the meta-analysis that SAS-085 designed and conducted on C2 agility.

The first is the generalized linear mixed model. A model can be linear or not and generalized or not. In the absence of knowledge about the type of relationship, the linear model is usually used. If the distribution of the measured variables is other than a normal one, a generalized model is more appropriate. Finally, the statistical model can be mixed or not which depends on whether the independent variable(s) is random or fixed. Fixed effects involve independent variables, or treatments, for which the only levels of interest are those included in the experiment as was the case when the treatment was one of five distinct C2 Approaches. In other situations, independent variables can take on a subset of an infinite number of possible values. In other words, controlled or observed values of a variable constitute a sample of a larger population of values. *Simulation model* (or *Experiment*) is the primary random variable in a meta-analysis. It represents a “sampling” of an infinite number of possible simulation models that maybe of interest to test. But for some reasons (e.g. they are unknown to the experimenter, they do not exist yet, or because it would be too costly to exploit them), the meta-analysis does not include them all. There is a still more important reason to considering *Simulation model* as a random variable. Random effect models deal with the heterogeneity[3] of the meta-analysis, an undesired property that occurs when simulation models differ on too many aspects. A method for dealing with this variability is explained in the next section. The *Endeavor space* is another example of a random variable in C2 Agility-related experiments.

[3] When there is more variation between the studies being included in a meta-analysis than what is expected by chance alone.

Meta-analyses are likely to combine both fixed and random effects in their design, requiring what is called a mixed model for their analysis. In such a model, *Simulation model* is defined as a block[4]. Blocks are groups of experimental units that are similar. By including blocking in a meta-analysis, the model captures the variability between and within blocks (simulation models) and can better estimate the impact of the fixed effects on the dependent variable(s). When experimental units are randomly assigned to a block, it is called randomized block design, a highly desired feature of an experiment. In the example measuring the impact of adopting a C2 Approaches on agility, it can be difficult to compare the average agility results for organizations that adopt two C2 Approaches if the agility values differ for a particular C2 Approach across the simulation models. A mixed model with *C2 Approach* as a fixed effect and *Simulation model* as a random effect will "subtract" any variability due to missing settings and the measures specific to each simulation model. It is possible for *Simulation model to* be considered a fixed effect even if it is not the treatment but, by doing so, findings are specific to the limited situations represented by the set of simulation models included in the meta-analysis. With *Simulation model* modelled as a random variable, findings apply to an infinite population of similar simulation models.

Multiple regression analysis is another useful tool. It estimates the relationship between one or more potentially explanatory variables, or predictors, on one dependent variable. The contribution of each predictor is calculated while keeping the other predictors constant. When using multiple regression, a meta-analysis must strongly consider including *Simulation model* as a predictor in order to take into account the blocking introduced in the experiment. This way, the effect of the S*imulation model* on the dependent variable is subtracted and the remaining effects are those that can be attributed to the other predictors.

3 SAS-085 Meta-Analysis

The SAS-085 NATO Research Task Group (RTG) on Command and Control (C2) Agility and Requisite Maturity was created with the objective of improving the understanding of the importance of C2 agility for North Atlantic Treaty Organization (NATO) and its member nations. Several papers present the results of C2 Agility-related case studies and individual experiments. However, each of these contributions was based upon a single experimental environment and/or simulation model. SAS-085, in order to produce more complete, robust, and generalizable set of findings undertook a meta-analysis of multiple simulation-based experiments. Specifically, SAS-085 members from five NATO member nations, namely USA, Portugal, Canada, United-Kingdom, and Italy jointly conceived a meta-analysis using multiple experimental platforms and simulation models. Some results are presented in a series of papers (Alberts, Bernier, Chan, & Manso, 2013; Bernier, Alberts, & Manso, 2013; Bernier, Chan, Alberts, & Pearce, 2013) that address between two and four hypotheses each. Some of those results are presented in this paper to support explanations. Conversely, the current paper provides background information to those papers by explaining the methodology and experimental setup.

3.1 Selecting Simulation Models and Developing Hypotheses

Five simulation models were initially known and considered by the SAS-085 experimentation team. A sixth simulation model was subsequently identified and determined to be applicable. These six simulation models all had been used in at least one independent experiment whose objectives were compatible with the objectives of the meta-analysis. The simulation models included in this meta-analysis are: IMAGE (Lizotte, Bernier, Mokhtari, & Boivin, 2013), WISE (Pearce, Robinson, & Wright, 2003), PANOPEA (Bruzzone, Tremori, & Merkuryev, 2011) and three variants of ELICIT (Chan, Cho, & Adali, 2012; Manso & Nunes, 2008; Ruddy, 2007).

[4] The term block takes it origin from the early ages of experimentation. Blocks where designated plots of land where various fertilizers or seeds where tested. Since plots may have had different intrinsic yields (e.g. better drainage), blocking allowed for subtracting the effect of the intrinsic yield of the plot from the total effect, leaving only the fertilizer or seed effect.

The formulation of the hypotheses for the meta-analysis generated considerable discussion and debate. The initial results of the analysis of the data generated caused the team to revisit the suitability of the measures employed and the formulation of the hypotheses themselves. One reason is that hypotheses are the interface between theory and the “hard” evidence as captured by the experimental data. There are, of course, multiple valid ways to test any hypothesis and the team sought to find the best approach given the available models and runs. Finally, contrary to words, the rigour and unambiguous language expressed by mathematical analyses leave far less room for interpretation. Unexpectedly, the SAS-085 team realized that even if the results of the meta-analysis were to prove erroneous, the process of conducting it would be extremely useful. Designing and conducting a meta-analysis fostered highly critical thinking and helped challenge assumptions. The reader is invited to consult the individual papers referenced to get a detailed description of the hypotheses tested.

3.2 Experimental Setup

Figure 1 illustrates a schema of the experimental design for the meta-analysis. There are two explicit and one implicit independent variables. The first independent variable, *C2 Approach*, can take on five different values (Conflicted, De-Conflicted, Coordinated, Collaborative, or Edge). An experiment instantiates anywhere from two to all five of the pre-defined C2 Approaches. The second independent variable, *Endeavor Space* represents a series of challenges within the operational or mission setting, each of which corresponds to a particular set of circumstances (CiCs) a collective may face. The set of experiment runs consists of C2 Approach / CiCs combinations so that each C2 Approach is employed in each circumstance. The endeavor space includes CiCs that involve various states of degraded and denied environments as well as other challenges that cause effects similar to those caused by a degraded environment (delays, increased work load). Finally, *Simulation Model*, or *Experiment*, is an implicit independent variable. It is of little interest in itself but is nevertheless captured because it represents a sample of a virtually infinite population of simulation models. As previously mentioned, using *Simulation Model* as a block of experimental units allows controlling for their difference and then reduces the variability that may hinder the effect of the C2 Approach. *Simulation Model* is a random effect, meaning that the findings from our six simulation models can be generalized to an infinitive hypothetical population of simulation models.

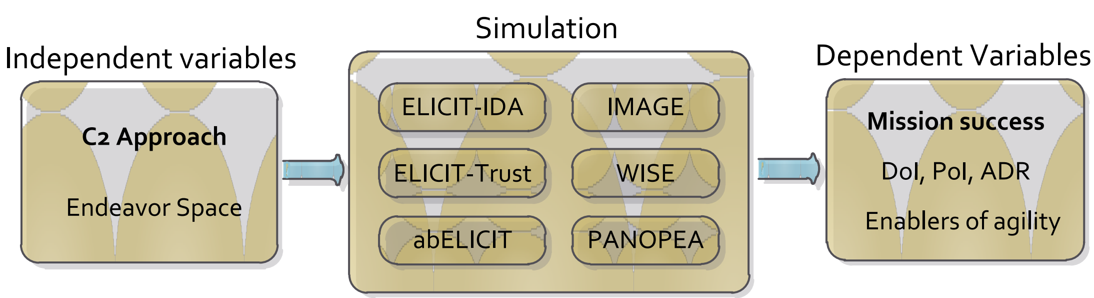

Figure 1: An example of experimental design for the meta-analysis.

3.3 Independent Variables

The first independent variable, *C2 Approach*, a fixed effect, is the treatment for this meta-analysis. It should be noted that not all of the simulation models implemented all of the C2 Approaches. The resulting design is thus non-balanced, i.e. values are missing for some combination of levels of *C2 Approach* and *Simulation Model*. For this reason, the average values of the outcome (dependent) variables such as *Agility score* were not computed as the arithmetic mean but instead as the least squares (LS) mean, or estimated marginal means. LS-means represent the mean response for each factor adjusted for the *Simulation Model* variable in the statistical model, including missing values. C2 Approaches designed across different simulation models were considered identical from a statistical analysis perspective (crossed design). The C2 Approaches instantiated in each of the six simulation models are identified in Table 1. Although each implementation of the C2 Approaches is different, verifications were conducted to ensure that they were equivalent as much as possible across all simulation models and all complied with the NATO NEC C2 Maturity Model.

Table 1: C2 Approaches implemented in each experiment.

	ELICIT-IDA (USA)	ELICIT-TRUST (USA)	abELICIT (Portugal)	IMAGE (Canada)	WISE (UK)	PANOPEA (Italy)
Conflicted		X		X		
De-Conflicted	X	X		X	X	X
Coordinated	X	X	X	X		
Collaborative	X	X	X	X	X	X
Edge	X	X	X			X

The primary role of the endeavor space is to deduce agility, i.e. the proportion of the endeavor space where a collective is successful. But endeavor space serves two additional purposes. First, the endeavor space corresponds to what is called a *noise factor* in the literature (Steinberg & Bursztyn, 1998). Such factors aim at recreating the natural variability found in the real-world and then at improving the external validity and robustness of the findings. Second, incorporating a large quantity of CiCs reduces the probability of selecting only CiCs that would be systematically detrimental or beneficial to some C2 Approaches (law of large numbers). A different endeavor spaces was defined for each experiment. The endeavor space of the meta-analysis was populated by combining all levels of all types of CiC for a given experiment (see Table 2). The endeavor space of all resulting experiments comprised 22 types of CiCs for a total of 231 instances of CiCs, far more than any individual experiment. CiC is a good example of where diversity must be sought. Nevertheless, a few CiCs were dropped because it would have taken too much time to simulate them all or because they were incompatible with other runs. Providing subcategories is another useful way to facilitate the verification of some independent variables.

Table 2: Endeavour space defined by the types of CiCs affecting the experiment-specific selves and their environment (with number of levels per type of CiC).

	ELICIT-IDA	ELICIT-TRUST	abELICIT	IMAGE	WISE	PANOPEA
Self	Network damage (3)	Message/Drop rates (3)	Infostructure degradation (2)	Latency (3)	Bandwidth efficiency (2)	
		Trust (3)	Agent performance (3)	Missing org (2)		Ship decision-making capability (2)
		Selfishness (3)	Organisation disruption (2)			Intelligence DM capability (2)
Environment	Challenge (4)		Key info. available (3)	Number of rebels (3)	Comm. link quality (2)	Number of pirates (2)
	Noise in information (3)			Crisis severity (3)		Weather condition (2)
	Cognitive complexity (3)					Misleading information (2)
#CiC	108	27	6	54	4	32

3.4 Dependent Variables

According to the NATO Network Enabled Operations (NEC) C2 Maturity Model (N2C2M2) developed by NATO SAS-065 and published by the DoD CCRP (Alberts et al., 2010), C2 Approaches differ on at least three major aspects: the allocation of decision rights (ADR), the pattern of interaction among entities (PoI), and, distribution of information among entities (DoI). Together they create the three dimensions that form the C2 Approach Space. An objective of the meta-analysis was to determine if C2 Approaches occupy different regions of the C2 Approach Space. The difficulty was to choose one or more proxies (metrics) of each dimension and then select one or more variables among those already captured by each experiment. In this case, a conceptual framework provided some guidance. Because of the large number of possible measures, it was decided that having diversified measures would capture more perspectives of the characteristics of these dimensions. The resulting Table 3, shows the definition of measures used in the meta-analysis to measure DoI, PoI, and ADR.

Table 3: Metrics for measuring the actual position in the C2 Approach Space

	DoI	PoI	ADR
ELICIT-IDA	Average percent of factoids received by each individual.	Scaled square root of number of information related transactions (post, pulls, shares).	Amount of individual with decision rights divided by total number of individuals.
ELICIT-TRUST	Average percent of factoids received by each individual.	Average number of links used.	Amount of individual with decision rights divided by total number of individuals.
IMAGE	Normalised difference between all variables values known by all individuals and the ground truth.	Sum of all co-conducted activities between organizations divided by the sun of all conducted activities.	Number of decisions allocated to the collective divided by the total number of possible decisions.
WISE	Mean HQ SA scores + (1-Eigenvector Centrality)).	Mean of the (normalised value of Sociometric status) + (1-Bavelas-Leavitt centrality) + Inverse path length + Clustering score / 4	1-Betweeness Centrality
PANOPEA	Average number successful received alerts against the total number of sent alerts.	Total number of communications among actors divided by number of alerts from intelligence	All the information taken directly by frigates and helos.

3.5 Modelling Effects and Examples of Results

An important hypothesis tested by the meta-analysis was that entities operating with more network-enabled C2 Approaches, like Collaborative and Edge, exhibit more agility. As previously stated, Agility is measured by the proportion of the endeavor space (CiCs) in which a collective is successful. This value is called the *Agility Score* and is calculated by averaging all values of *Mission Success* measured for all CiCs simulated for a given C2 Approach. Table 1 shows the agility scores calculated for each C2 Approach for every experiment (or simulation model). Each simulation model is different in term of the situation simulated (some might be more complicated or challenging), their implementations of the C2 Approaches, and the metrics and criteria used to calculate *Agility score*. A linear mixed model was used to test the hypothesis. The results obtained from the meta-analysis support the hypothesis that more network-enabled C2 Approaches are more agile (for details see Bernier et al. (2013)). For this discussion, it is important to note that the resulting average agility score for each C2 Approach is not the geometric mean (e.g. Conflicted Agility Score = (0.04 + 0.39)/2= 0.22 but the estimated marginal means (or least squares means). This is the reason that the agility scores of IMAGE, are higher not only for Conflicted, but are also higher for the other C2 Approaches as well. The statistical model used here “understands” that IMAGE is biased toward higher values. The mixed model removes this bias and then produces a lower agility score for conflicted, which is what we would have desired instinctively.

Table 4: Agility scores for each C2 Approach and experiments – least square means (M) and standard error (SE).

C2 Approach	ELICIT-IDA	ELICIT-TRUST	abELICIT	IMAGE	WISE	PANOPEA	LS-Mean
Conflicted		0.04		0.39			0.09 (0.10)
De-Conflicted	0.06	0.06		0.50	0.21	0.41	0.14 (0.09)
Coordinated	0.10	0.06	0.02	0.54			0.20 (0.09)
Collaborative	0.26	0.18	0.13	0.89	0.42	0.72	0.39 (0.09)
Edge	0.55	0.46	0.33				0.59 (0.09)

Trying to find the average position in the C2 Approach Space for each of the C2 Approaches provides another example of how this statistical model works. The values of DoI, PoI and ADR were calculated for each CiC for every experiment and C2 Approach (see Figure 2). It is obvious that values are grouped differently for the different experiments. The linear mixed model takes into account these differences. For instance, it may be difficult by visual inspection to declare that Collaborative has higher value of DoI than Coordinated. And assuming that these values are randomly distributed does not help, for the result of the statistical test is likely to be non-significant. By using a mixed model modeling *C2 Approach* as a fixed effect and especially *Simulation model* as a random effect, the differences between the simulation models were "subtracted". And then yes, the difference was statistically significant. The reader is invited to consult Bernier, Chan et al. (2013) for more details.

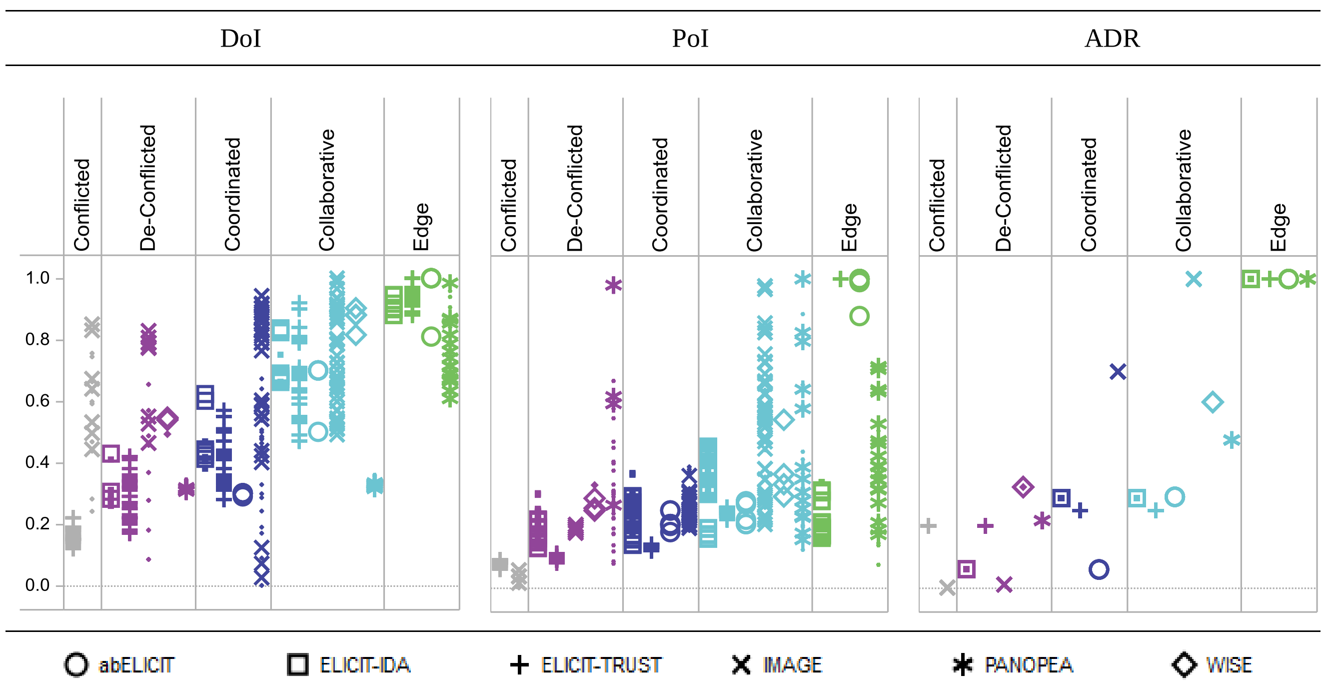

Figure 2: Mapping of all CiCs into each axis of the C2 Approach Space.

Finally, even though the values of DoI, PoI, and ADR where bounded between 0 and 1 for all experiments, the result of the analyses produces estimated marginal means over and above zero for ADR as illustrated in Table 5. These unexpected results are due to the linear estimation of effects used by the statistical model.

Table 5: Average values in the C2 Approach Space of all CiCs tested under each C2 Approach – estimated marginal means (standard error).

C2 Approach	DoI	PoI	ADR
Conflicted	0.36 (0.12)	0.04 (0.07)	-0.05 (0.13)
De-Conflicted	0.41 (0.11)	0.25 (0.06)	0.10 (0.12)
Coordinated	0.43 (0.11)	0.28 (0.06)	0.41 (0.12)
Collaborative	0.63 (0.11)	0.43 (0.06)	0.50 (0.12)
Edge	0.98 (0.12)	0.44 (0.06)	1.08 (0.12)

4 Conclusion

This paper presented a methodology for designing and conducting meta-analyses involving many simulation models and research teams. This paper provides guidance for applying the principles of meta-analysis to the context of simulation-based experiments. The most useful concepts to be applied and notable differences were highlighted, including the drawbacks and the benefits of various options to design the experiments. Finally, this paper illustrated a few steps of the design process with the international SAS-085 meta-analysis.

As the pool of simulation models reaches a significant size, there is growing potential for applying the methodology explained in this paper. Many improvements are definitely possible. Statistical analysis and experimental design are complex fields and it is likely that better methods exist and were not introduced in this paper. Nevertheless, the method of prospective meta-analysis conducted on multiple experiments explained here provides a number of benefits when compared to conducting separate experiments or waiting for more experiments to be completed before conducting a retrospective meta-analysis. In summary, although there are many challenges to overcome with combining multiple experiments/simulation models in a meta-analysis, the benefits should exceed the drawbacks.

Acknowledgements

The authors gratefully acknowledge the inputs from Agostino Bruzzone, Kevin Chang, Marco Manso, and Alberto Tremori for their very valuable contributions to this experiment with simulation data and comments.

References

Alberts, D. S., Bernier, F., Chan, K., & Manso, M. (2013). C2 Approaches: Looking for the “Sweet Spot.” *Proceedings of the 18th ICCRTS*. Alexandria, VA, USA.

Alberts, D. S., Huber, R. K., & Moffat, J. (2010). *NATO NEC C2 maturity model*. United-States: DoD Command and Control Research Program.

Barton, R. R. (2004). Designing simulation experiments. *Proceedings of the 36th conference on Winter simulation* (pp. 73–79).

Bernier, F., Alberts, D. S., & Manso, M. (2013). International Multi-Experimentation Analysis on C2 Agility. *Proceedings of the 18th ICCRTS*. Alexandria, VA, USA.

Bernier, F., Chan, K., Alberts, D. S., & Pearce, P. (2013). Coping with Degraded or Denied Environments in the C2 Approach Space. *Proceedings of the 18th ICCRTS*. Alexandria, VA, USA.

Bravata, D. M., & Olkin, I. (2001). Simple Pooling versus Combining in Meta-Analysis. *Evaluation & the Health Professions*, *24*(2), 218–230.

Bruzzone, A. G., Tremori, A., & Merkuryev, Y. (2011). Asymmetric Marine Warfare: PANOPEA a Piracy Simulator for Investigating New C2 Solutions. *Proceeding of the SCM MEMTS Conference* (p. 32). Saint-Petersburg, Russia.

Chan, K., Cho, J. H., & Adali, S. (2012). Composite Trust Model for an Information Sharing Scenario. *Proceedings of the Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC)* (pp. 439–446). Fukuoka, Japan.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100.

Egger, M., & Smith, G. D. (1998). Meta-analysis bias in location and selection of studies. *Bmj*, *316*(7124), 61–66.

Ghersi, D., Berlin, J., & Askie, L. (2011). Prospective meta-analysis. *Cochrane Handbook for Systematic Reviews of Interventions*, *5*, 1.

Humphreys, P. (2004). *Extending Ourselves:Computational Science, Empiricism, and Scientific Method: Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT sloan management review*, *52*(2), 21–32.

Lizotte, M., Bernier, F., Mokhtari, M., & Boivin, E. (2013). *IMAGE Final Report: An Interactive Computer-aided Cognition Capability for C4ISR Complexity Discovery* (No. TR 2013-397). Québec, Canada: Defence R&D Canada - Valcartier.

Manso, M., & Nunes, P. (2008). ELICIT and the Future C2: Theoretical Foundations for the Analysis of ELICIT Experiments. *Proceedings of the 13th ICCRTS* (pp. 17–19). Seattle,WA, USA.

Pearce, P., Robinson, A., & Wright, S. (2003). The Wargame Infrastructure and Simulation Environment (Wise). *Proceedings of the Knowledge-Based Intelligent Information and Engineering Systems Conference* (pp. 714–722). Oxford, UK.

Ruddy, M. (2007). ELICIT–The Experimental Laboratory for Investigating Collaboration, Information-sharing and Trust. *Proceedings of the 12th International Command and Control Research and Technology Symposium* (pp. 19–21).

Sargent, R. G. (1994). Verification and validation of simulation models. *Proceedings of the 26th conference on Winter simulation*, WSC '94 (pp. 77–87). San Diego, CA, USA: Society for Computer Simulation International.

Steinberg, D. M., & Bursztyn, D. (1998). Noise factors, dispersion effects, and robust design. *Statistica Sinica*, *8*, 67–86.

Sterne, J. A. C., Egger, M., & Smith, G. D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, *323*(7304), 101–105.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075.