Pragmatic Expert Elicitation for Defence Capability Analysis

A. Donohoo[a], J. Chisholm[a] and J. Barles[a]

[a]Maritime Capability Analysis
Joint and Operations Analysis Division
Defence Science and Technology Group
Email: andrew.donohoo@dst.defence.gov.au

Abstract: Defence Capability Analysis examines the ability of a real or notional force to conduct military tasks. When faced with a paucity of quantitative data or lack of sufficient models, there is a need to rely on Expert Elicitation as a pragmatic form of preliminary assessment. Typically, Subject Matter Experts (SMEs) are brought together to work through a hierarchy of questions about the capability area under investigation.

In these activities, it is essential that individual responses are captured adequately, along with positive and negative feedback collected from the wider group, because analysts need to comprehend the substance of discussions and to identify where and why agreement or otherwise occurred. An overall sense of progress via aggregation of responses is needed by participants to frame the ongoing debate. Facilitators benefit from real-time support to guide discussion purposefully. Stakeholders wishing to explore the data and analysis products mean these must be persisted and remain accessible.

The analysis requirements have led the authors to implement a prototype web application called the Maritime Expert Elicitation Capability Analysis Tool (MEECAT). An overview of the system architecture is given, and the main strengths and weaknesses are discussed. MEECAT features include hierarchical structuring, anonymity of respondents, and a self-organising feedback loop in which participants comment on and promote/demote the responses of others. The tool assists in group facilitation through the visual way responses are aggregated, so that consensus or high variance are easily identified, and overall progress displayed. Participants, facilitators, analysts and interested observers benefit from the immediate remote browsability of all responses.

Some suggestions for enhancement of the tool and analytical technique are offered, based on experience, user comments and the enduring analytical aims driving the tool's inception.

The validity of this approach is ultimately gauged by its utility in systematically capturing warfighting problems and enumerating and articulating possible causes. MEECAT has been adopted successfully for a range of client centric activities from current force warfighting capability assessments to high level capability requirements elicitation and Navy strategic planning.

Keywords: *Capability analysis, experimentation, assessment frameworks, expert elicitation, structured analysis, web applications*

1. INTRODUCTION

Maritime Capability Analysis is concerned with understanding the ability of a force to deliver military effects, with a view to informing military decisionmakers about the changes that are needed, given the potentially complex interplay of numerous causal factors, to maximise the likelihood of success in achieving these goals.

From its origins [Gass et al. (2005)], Military Operations Research (MOR) has embraced both quantitative and qualitative problem solving techniques, choosing the most appropriate given the nature of the problem at hand and the time and resources available, sometimes modifying well-used methods in the search for improvement. The application of quantitative techniques for Military Operations Research gave rise to its success (see Jaiswal (2012)). Some MOR problems benefit from the use of closed-form mathematical models, whilst for other problems the use of computational techniques as in Law et al. (1991) allows the numerical response of modelled systems to be mapped across their state-space and over time.

However, for a variety of reasons modelling or simulation may not be suitable. Overall capability is sensitive to factors that go beyond mere platform system performance to include many non-materiel issues. The degree of complexity introduced by the human element can defy credible systematic representation [Sargent (2013)]. The systems under study may be ill-defined, insofar as their physical properties and behaviour are unknown or unknowable. The relevance of hypothesised capability models can be called into question, especially where new concepts for warfighting must be explored, tested and refined before they can be encoded with any confidence. This is exacerbated by the frequent paucity of real world quantitative data for performing and/or validating detailed analysis, forcing analysts to rely heavily on wide ranging parameterisations and simplifying assumptions.

One way of obtaining informed responses to emergent problems is to draw upon the guidance and judgment of experts in the domain under study. An empirical-deductive process of experimentation involving real human agents can identify underlying cause-and-effect relationships in complex military capabilities [Bowley et al. (2006, Kass (2006)]. But the requisite amount of time and resources for a thorough experimentation campaign are not always available.

Within the Maritime warfare domain there is an increasing need to keep senior Navy decision makers informed about changes in high level capabilities of the Navy. Periodic assessments of current high-end warfighting capabilities have been sought to discover gaps and prioritise areas for remediation. Other studies have involved the assessment of future naval capability for preliminary needs analysis in acquisition projects, or have sought to capture long term capability risks in order to refine strategic plans. The focus has widened from assessing individual systems in a very controlled setting, to assessing the effectiveness of a larger scale task force against simultaneous threats in undersea, surface and above surface domains. The scope includes tactical and operational and strategic level effects, for example ranging from Local Anti-Submarine Warfare (defending a ship against torpedo attack) to Force ASW (detect and deter submarines over a geographic region using ships, submarines and aircraft) to Theatre ASW (stopping enemy submarines from leaving their bases). The context is not only the current force facing a known threat using established doctrine, but also a notional future force facing a predicted threat whilst applying developmental warfighting concepts.

This inherent complexity and uncertainty along with the need for timeliness and repetition has led to an awareness of the pragmatic value of inductive approaches which rely on Expert Elicitation. These can serve as a precursor to more focused quantitative studies, where the opinions provided by the subjective experience and observational history of Subject Matter Experts (SME), can be tested for validity by objective analyses. The value of this approach is gauged by its utility in systematically capturing warfighting problems and enumerating and articulating possible cause and effect relationships rather than its power to isolate specific causes or make predictions about capabilities that can solve warfighting problems [Rubel (2006)].

1.1. Analysis Framework

The technique involves the application of a standardized Force Assessment Framework developed by Chisholm (2015) based on the US Department of Defense Architecture Framework (DoDAF) (2010). Selected Maritime Operations Tasks (MOTs) from a hierarchy are combined to form the Capabilities that produce specific military Effects. A given Capability can be broken down into Tasks as shown in **Figure 1**. The MOTs have been refined via consultation with stakeholders, reference to published Naval doctrine [RAN (2000)] and comparison to extant defence architecture standards [NWDC (2000)]. These MOTs are the invariant frame of reference upon which multiple capability assessments can be compared.

The sets of Capabilities and Tasks chosen for a given assessment depend on the military Effects to be achieved in a situating warfighting scenario, including timeframe, threats and force options under investigation. Further resolution to the analysis is usually provided by including in the assessment several cross-cutting dimensions as found in recognised defence capability management frameworks[1], which allows non-materiel capability gaps to be addressed during enquiry. These dimensions are Tactics, Techniques and Procedures, Organisation, Training, Equipment, Manning (TOTEM). The resulting question framework typically follows the hierarchy of Scenario → Force Option → Effect →Task → TOTEM category. A system for the collection of SME assessments using this framework is described in detail below.

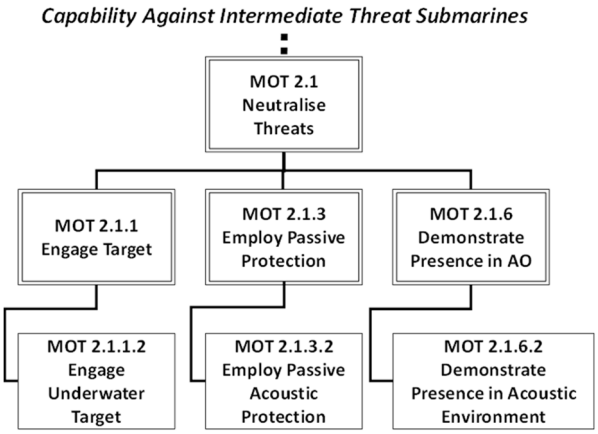

Figure 1. A top level *Neutralise Threats* Task and children. *Capability Against Intermediate Threat Submarines* includes other tasks not shown here.

1.2. Assessment Process

The structured assessment process aims to highlight capability issues and to understand the range of viewpoints and reasons for agreement or disagreement through focused SME discussion and feedback. To date this has occurred by co-locating SMEs during a one or two-day workshop at a secure location at either DST or client facilities. A skilled facilitator plays a role in guiding the group through the list of Tasks one by one at a steady rate. The facilitator may make a decision to skip one or more Tasks. Each Task is completed in three phases.

The first phase is open discussion and helps the group to interact to generate ideas and stimulate individual thinking. This loosely follows commonly used brainstorming methods such as Lunenburg (2011). The facilitator situates the group in the correct warfighting context (Scenario/Force Option/Effect), and steers the discussions to explore the issues or clarify different viewpoints as needed.

In the second 'survey' phase, individual responses of SMEs are captured. Because of the likelihood of competing as well as unvoiced opinions in these group activities it is essential that all responses are captured privately and anonymously. It is customary to obtain a decision response value for each question (using a five-point Likert Scale), and also to capture a free text rationale behind each person's decision.

To provide contestability the participants privately evaluate others' arguments and give anonymous positive or negative feedback through comments and/or votes. This occurs during the third phase. This critiquing step is like a single iteration of the Delphi Method [Brown (1968)]. Whilst many applications of the Delphi technique seek convergence of group opinion by iterating through several rounds, here diverse opinion provides analysts with insight into the capability assessment's substantive issues.

Following completion of the data collection process, analysts work through the numerous comments to identify the major themes, points of concurrence and residual conflict. Reporting categorises the key issues that have emerged (e.g. capability gaps and their perceived causes) and provides a high level statistical summary of response scores to compare to results from other collection exercises.

2. MEECAT APPLICATION

2.1. Software Architecture Overview

The analysis concept has led to a prototype software implementation called the Maritime Expert Elicitation Capability Analysis Tool (MEECAT). Some screen shots are shown in **Figure 2**. The MEECAT web application was developed in PHP to extend the survey functionality provided out-of-the-box by the popular open source web application Limesurvey [Schmitz (2012)]. All code can be hosted on a DST managed web server. The customised extensions access the MySQL database used by Limesurvey.

[1] The US military use the capability management dimensions of Doctrine, Organizations, Training, Materiel, Leader Development, Personnel, Facilities ("DOTMLPF"). The UK Ministry of Defence uses the categories of Training, Equipment, Personnel, Information, Concepts and Doctrine, Organisation, Infrastructure, Logistics (aka "TEPID OIL"). The Australian Defence Organisation uses Fundamental Inputs to Capability (FIC) which consist of Command and Management, Organisation, Major Systems, Personnel, Supplies, Support, Facilities, Collective Training, Industry.

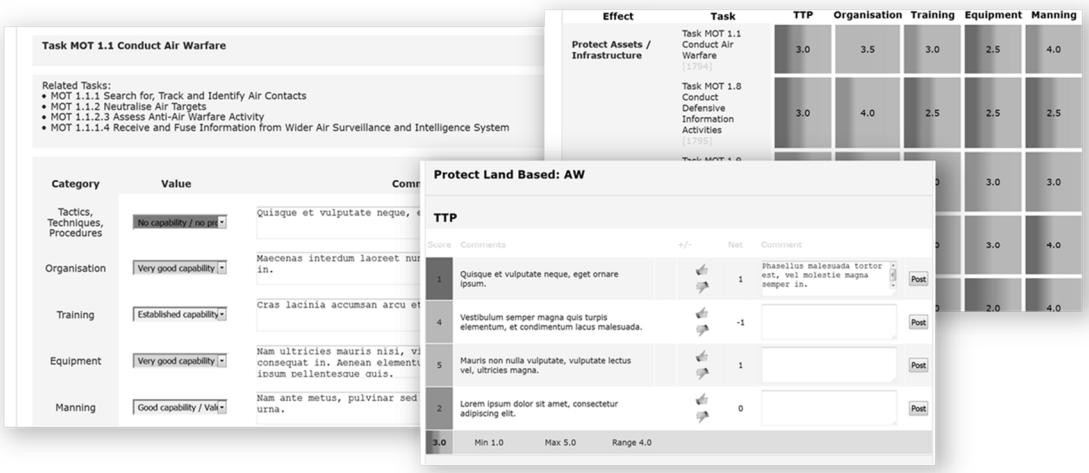

Figure 2. Dummy MEECAT web forms for TOTEM assessments of a Task (rear left) and for voting and feedback on whole of group assessment responses (front right). A scorecard page (rear right) shows the aggregation of group responses for the Tasks under the Effect of *Protect Assets / Infrastructure.* The colour gradient in scorecard cells indicates the relative distribution of scores across the five response levels.

Capability assessment workshops prior to the development of MEECAT used a combination of email, spreadsheets, and Microsoft Access, and suffered from a range of problems such as unnecessary duplication of setup effort, low operational concurrency, poor maintainability, fragility and proliferation of files across the multiple client machines making data aggregation slow and cumbersome. With MEECAT, client sessions are contained in a web browser and all data is persisted robustly to a single database with a known schema which can be queried and updated remotely and independently using a variety of tools. The calculation and display of group-wide metrics and summaries back to the users is instant and on demand. The numerous formatted questions required to setup for each new workshop can be generated automatically and stored in the database using a set of pre-written scripts instead of being crafted individually by hand. To date, survey automation as well as tailored post-collection analysis scripts have been written using PHP, Python and R languages.

Because no code resides on the client machines, MEECAT is compatible with the use of thin client hardware that is common on classified networks which permits activities to be run at any number of Defence establishments with minimal setup requirements. Stakeholders at other physical sites can access live analysis results either during or after the collection event. SMEs can participate without being physically co-located with the web server or with each other.

After authenticating to MEECAT the user is able to access the survey pages (to complete Phase 2 of the assessment process), voting pages (for Phase 3) or a home page which shows the overall progress via a scorecard as shown in **Figure 2**. Users drill down to address specific questions and provide commentary on collected responses using hyperlinks beginning from each cell.

2.2. MEECAT Strengths, Weaknesses and Recommendations

The following appraisal of the tool and analytical technique is informed by literature including Groves et al. (2011), as well as practical advice collected from analyst and user comments over the product lifetime.

The use of expert judgements means that the data being collected relies on a distillation of respondents' accumulated experience and resulting views. It is conceivable that the recollections and impressions of SMEs vary from day to day and can change in group situations due to the influence of others. The judgments expressed are therefore unverifiable and do not necessarily produce the same result as would be obtained if detailed objective measurements were possible. Further research into efficient ways of capturing knowledge and providing supporting evidence would lead to an improved process. A good overview of starting alternatives is given in Richards et al. (2010).

For now, the assessment system delivers value by providing a systematic way of focusing on contextualised topic areas (eg Effects, Tasks, TOTEMs) with a combination of individual and group responses, allowing successive refinement as the initial check for validity. Users are encouraged to populate the justification comment field for each of their assessments, drawing on their experience and making their 'evidence', perspectives and reasons available to others.

The use of a standardised framework should help compare results from one participant to another, one data collection activity to another, and across longitudinal studies. One of the continuing points of inconsistency and ambiguity for respondents is the understanding of TOTEM maturity levels used in the Likert scale, which are hard to differentiate between meaningfully. Whilst clearer definitions and explanations of response options are desirable, they are inherently ill-defined constructs without standard meanings so are open to widely different interpretations. This has the effect of introducing measurement error, increasing variance and decreasing confidence in the results.

Sufficient diversity of respondents is one of the quality factors to consider so it is beneficial to impress this need upon the client. The SMEs should be drawn from outside the immediate client workgroup otherwise the demarcation lines between enquiring and reporting are blurred. Ideally the group of SMEs should provide equal coverage of all the TOTEM dimensions as well as the various capability areas, but in practice it is often difficult to obtain this balance, whether due to unavailability, varied seniority, or stove-piping of expertise. This is one reason why the option of 'Don't know' is offered as a response. The population from which experts can be drawn is small thus most activities have a sample size of fewer than 20 participants. This conspires to make any analysis of numerical results low in statistical power and at risk of bias. Whilst the data capture system is designed to scale to much larger sample sizes than currently used it could be supplemented by capturing demographic information to aid with statistical analysis and stratification.

A key part of a successful analysis is ensuring that the warfighting context is relevant and sufficiently detailed and that there is a suitable trade-off between coverage and conciseness. To ensure SMEs have effectively understood the Effects and Tasks in the context of the scenario, it has proven helpful to include descriptions of key terms and associated definitions, in the form of HTML links or popups which are instantly accessible throughout the data collection period. One untested alternative is to use a ranking scheme such as Cameron et al. (2010) in which users prioritise selections from a subset of alternatives. This could have the advantage of reducing workload, for example by asking users to rank their top most-concerning TOTEM categories for a Task instead of scoring all TOTEMS individually according to a Likert scale. The collection of redundant information would also be reduced.

In the aggregation of responses, variance is a useful measure, and the level of uncertainty should be reflected in the way results are presented to clients. One way that has been prototyped is making the most uncertain elements in the facilitator scorecard of lower colour intensity. More importantly, communication of the results should not be reduced to simplistic numerical scores which hide the insights achieved from the analysis of textual content and group discussions. Thus it is essential for analysis to take several different approaches to collection of data including categorical scores, voting, trending patterns in topic responses, as well as more meaningful but laborious textual analysis.

One of the chief areas of concern is finding balance between resolution of questions and respondent workload. Whilst the Effect to Task breakdown provides sufficient resolution it does lead to some repetition. The trade-off is currently managed by the schedule and pace that is set by the facilitator. The workshop designers too must take into account the increasing fatigue as the data collection process continues, and the effect that it has on the quality, length and number of responses. For respondents who type more slowly than others, data entry is an added impediment to expression. Although the systematic approach to data collection that is used is desirable from an analyst's perspective, it can admittedly become tedious for the respondents to answer many questions of the same form and their ability to focus and make consistent judgements is affected.

Several alternatives could be tested. One option is to remove the requirement for all respondents to answer every question with the same level of detail. For instance, each could address just the areas in which they have the most specialised knowledge. Another option is to use a quick first pass, filling out just the scores, to identify areas for follow-up with more detailed textual capture. Where voting is required, each respondent could be randomly given a small subsample of user responses to process, so long as each contributed idea received the same number of views to maintain uniform representation.

With minor modifications MEECAT could be used for slow-time online-only discussions but this has not yet been trialled. As group size increases above about a dozen, the current approach used by MEECAT produces an overabundance of comments, many of which are similar. Ideally the ideas displayed in the top level view are all different, minimising redundancy, and supporting arguments can be added or viewed if a user drills down. Methods that emphasise the main arguments, or allow progressive self-organisation and promotion of ideas on merit by crowd based filtering (e.g. Klein et al. (2015), Anderson et al. (2012)) or by the weighted reputation value of respondents (as in Marsh et al. (2013)) avoid some of the shortcomings of the current process. For larger-scale online systems, responses are typically gathered over long timeframes and by default a user is exposed to just a few of the top rated previous entries. Whilst this reduces reading effort it can

contribute to bias as earliest posted comments receive more views, and already popular comments are attractors for like feedback.

The respondents aren't the only participants to suffer from question fatigue. The facilitator is faced with numerous and significant cognitive challenges including directing conversation and response flow, simultaneously reading and identifying and understanding key points, whilst managing time. Much hinges on the skill of the facilitator so assistance via software is all the more important. By displaying a 'scorecard' covering the whole question hierarchy the trends can be observed and the facilitator can tease out the reasons for strong consensus on unusually high or low scores or the reasons for strong differences of opinion where that occurs. With the capture and live display of sorted voting results MEECAT helps facilitators and analysts visualise contentious and consensus topics, see which responses are most popular or unpopular, and see which have attracted the most response activity.

Nevertheless it is advantageous for the facilitator to be well versed in the capability areas under investigation and have sufficient prior knowledge of the likely 'hot topics' that ignite discussion. The facilitator can still be swamped by the amount and pace of textual responses for each question, which is exacerbated as the sample size increases unless filtering of the most relevant ideas is possible. For closed group workshops it may be preferable if someone on the client's staff facilitates or co-facilitates, and analysts focus on managing the data collection and interpretation process. This also helps with ownership and engagement of the client. Having more than one facilitator take turns helps create variety and maintain freshness.

The limitations of the quantitative value in the technique should be understood. The scorecard presents a descriptive statistical summary, serving as a concise way of showing stakeholders the sentiment of the group and highlighting strengths and deficiencies in capability characteristic of particular TOTEM dimension and/or Effect and/or Task. Whilst a range of statistics can be chosen via URL parameters, the median is preferred for showing the central tendency of the scores (**Figure 2**), because parametric statistics such as the mean are inappropriate for ordinal data. The colour coded frequency distributions readily show scoring splits for a given cell. The large number of survey questions and the degree of variance in responses usually means that there is a dominant overall similarity of appearance from one scorecard cell to another, which masks visualisation of the 'unusual' events. Where more thorough statistical processing would be useful is in determining the significance of apparent trends across TOTEM categories or Tasks, and making the most relevant findings more distinctive. Providing users with more control over visualisation, such as customisation based on alternative metrics and statistical tests, would be advantageous.

Alternative statistical techniques for analysis of ordinal values (see Agresti (2010)) have not been considered a high priority to date. Apart from the metrics related to capability scores, there are other analytical techniques that are interesting to analysts such as the correlation of scoring patterns among respondents, the analysis of social networks apparent within the feedback and voting phases, and the alternative ways of sorting responses based on textual content, sentiment or vote distribution. Most recently numerical analysis of response data has investigated translating the five level scores used in MEECAT to positive or negative sentiment values, to quantify and account for the effect of scorers who have extreme views compared to others.

3. CONCLUSION

MEECAT has demonstrated its utility for a range of facilitated client centric activities from Campaign Assessments (current force warfighting capability assessments) to workshops for high level Operations Analysis requirements capture. A version of the tool has been in use for over four years.

MEECAT supports the Maritime Force Assessment Framework by providing characteristics to aid systematic and efficient data capture such as hierarchical structuring, anonymity of respondents, a self-organising feedback loop in which participants comment on and promote/demote the responses of others. The tool assists in group facilitation through the visual way responses are aggregated, so that consensus or high variance are easily identified, and overall progress displayed. Participants, facilitators, analysts and interested observers benefit from the immediate browsability of all responses.

As previously stated the greatest analytical value comes from the comments provided by respondents. The scores on their own do not provide stakeholders with actionable information. Future developments in MEECAT should do a better job of efficiently capturing and displaying the most relevant ideas, issues and themes and the relationships between these. This would not only improve client reporting timelines, it would also enable all interested stakeholders to independently explore and discover patterns hitherto unseen.

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, John Wiley & Sons.

Anderson, A., D. Huttenlocher, J. Kleinberg and J. Leskovec (2012). Discovering value from community activity on focused question answering sites: a case study of stack overflow. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. Beijing, China, ACM: 850-858.

Bowley, D., P. Comeau, R. Edwards, P. J. Hiniker, G. Howes, R. A. Kass, P. Labbé, C. Morris, R. Nunes-Vaz and J. Vaughan (2006). *Guide For Understanding and Implementing Defense Experimentation (Guidex)-Version 1.1*, The Technical Cooperation Program (TTCP).

Brown, B. B. (1968). Delphi process: A methodology used for the elicitation of opinions of experts, DTIC Document.

Cameron, F. and G. Pond (2010). Military Decision Making Using Schools of Thought Analysis–A Soft Operational Research Technique, with Numbers. Archive of the 27 International Symposium on Military Operational Research (OR Society Defence SIG), see: http://ismor.cds.cranfield.ac.uk/27th-symposium-2010.

Chisholm, J. (2015). *The Application of the Australian Defence Architecture Framework to Maritime Force Assessment*. DSTO Technical Report, DSTO.

Naval Warfare Development Command. (2000). *Naval Mission Essential Task List (NMETL) Development Handbook*.

Gass, S. I. and A. A. Assad (2005). *An Annotated Timeline of Operations Research: An Informal History*, Springer Science & Business Media.

Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau (2011). *Survey Methodology*, John Wiley & Sons.

Jaiswal, N. K. (2012). *Military Operations Research: Quantitative Decision Making*, Springer Science & Business Media.

Kass, R. A. (2006). *The Logic Of Warfighting Experiments*, Assistant Secretary of Defense (C3I/Command Control Research Program) Washington DC.

Klein, M. and A. C. B. Garcia (2015). High-speed idea filtering with the bag of lemons, *Decision Support Systems* 78: 39-50.

Law, A. M., W. D. Kelton and W. D. Kelton (1991). *Simulation Modeling and Analysis*, McGraw-Hill New York.

Lunenburg, F. C. (2011). Decision making in organizations, *International Journal Of Management, Business and Administration* 15(1): 1-9.

Marsh, B. D. and N. C. Gloy (2013). Managing content based on reputation, Google Patents.

Royal Australian Navy (2000). *Australian Maritime Doctrine*.

Richards, H. J. and R. H. Pherson (2010). *Structured Analytic Techniques for Intelligence Analysis*, CQ Press.

Rubel, R. C. (2006). *The Epistemology of War Gaming*, Naval War College, Newport RI.

Sargent, R. G. (2013). An introduction to verification and validation of simulation models. Winter Simulation Conference (WSC), 2013 IEEE.

Schmitz, C. (2012). LimeSurvey: An open source survey tool. LimeSurvey Project Hamburg, Germany. URL http://www.limesurvey.org.

US Department of Defense (2010). *US Department of Defense Architecture Framework Version 2.02*.