



Generalising from Novices to Experts in Military Studies

The work described in this document has been undertaken by the Human Factors Integration Defence Technology Centre, part funded by the Human Capability Domain of the U.K. Ministry of Defence Scientific Research Programme.

© Human Factors Integration Defence Technology Centre 2007. The authors of this report have asserted their moral rights under the Copyright, Designs and Patents act, 1988, to be identified as the authors of this work.

ReferenceHFIDTC/2/WP1.2.5/2

Version.....3

Date..... 11 July 2007

Authors

G. Walker	Brunel University
N. Stanton	Brunel University
Laura Rafferty	Brunel University
Darshna Ladva	Brunel University
D. Jenkins	Brunel University
P. Salmon	Brunel University

Contents

1	Executive Summary	1
2	Introduction	2
2.1	Aims and Objectives	2
2.2	Expertise	3
2.3	Experts and Novices	6
3	Methodology	8
3.1	Design	8
3.2	Participants.....	9
3.3	Materials.....	9
	3.3.1 Participant Instructions.....	9
	3.3.2 Dependant Measures	11
3.4	Procedure.....	11
	3.4.1 Phase 1 – Pre-Briefing.....	11
	3.4.2 Phase 2 – The Combat Estimate.....	11
	3.4.3 Phase 3 - Questionnaires	12
4	Results	13
4.1	Time	13
4.2	Situational Awareness.....	15
	4.2.1 Accuracy	15
	4.2.2 Knowledge Networks	17
	4.2.2.1.1 Stage 1	17
	4.2.2.1.2 Stage 2.....	18
	4.2.3 Knowledge Quantity.....	21
	4.2.4 Knowledge Type	22
	4.2.5 Interconnectivity of Knowledge	23
4.3	Mental Workload	23
4.4	Proving the Null Hypothesis.....	26
5	Conclusions.....	29

6	References and Bibliography	31
---	-----------------------------------	----

List of Figures

FIGURE 1 – RASMUSSEN’S DECISION LADDER. NOVICE PERFORMANCE WOULD PROCEED UP AND DOWN THE FULL HEIGHT/LENGTH OF THE DECISION MAKING PROCESSES. EXPERT PERFORMANCE (SHOWN) RELIES ON SHORTCUTS.	6
FIGURE 2 – REPRESENTATION OF THE URBAN BATTLESPACE AREA.	10
FIGURE 3 – BAR CHART ILLUSTRATING THE MEDIAN TOTAL TIME TAKEN BY THE EXPERT AND NOVICE GROUPS.	14
FIGURE 4 – THE ACCURACY ACHIEVED ACROSS ALL SEVEN QUESTIONS SUMMED TO GIVE AN OVERALL MEDIAN ACCURACY FOR EXPERT AND NOVICE GROUPS.	15
FIGURE 5 – LINE GRAPH ILLUSTRATING EXPERT AND NOVICE MEDIAN ACCURACY (%) ACROSS THE DIFFERENT PHASES OF THE COMBAT ESTIMATE.	17
FIGURE 6 – WORD FREQUENCY GRAPH ILLUSTRATING THE MANNER IN WHICH THE SCREE (THE SITUATION SPECIFIC WORDS) WERE REMOVED FROM THE ANALYSIS TO LEAVE ONLY THOSE ELEMENTS THAT OCCURRED CONSISTENTLY.	18
FIGURE 7 – KNOWLEDGE NETWORK FOR THE NOVICE GROUP	19
FIGURE 8 – KNOWLEDGE NETWORK FOR THE EXPERT GROUP	20
FIGURE 9 – BAR GRAPH SHOWING THE MEDIAN (AND 95% CONFIDENCE INTERVAL) FOR KNOWLEDGE QUANTITY.	21
FIGURE 10 – BAR CHART SHOWING OVERALL WORKLOAD FOR THE NOVICE AND EXPERT GROUPS.	24
FIGURE 11 – LINE GRAPH SHOWING THE RESULTS OBTAINED WHEN THE ANCHOR AND TREATMENT APPLICATIONS OF THE NASA TLX QUESTIONNAIRE WERE SUBTRACTED FROM EACH OTHER FOR BOTH EXPERT AND NOVICE PARTICIPANTS.	25

List of Tables

TABLE 1 – COMPARISON BETWEEN EXPERTS AND NOVICES IN TERMS OF THE TYPE OF SITUATIONAL ELEMENTS EXTRACTED FROM THE SCENARIO SHOWN WITH THEIR ATTENDANT POSITIONAL CENTRALITY SCORES (IN RANK ORDER: MOST CENTRAL ELEMENTS FIRST).	22
TABLE 2 – RELATIVE CONTRIBUTION(S) THAT EACH SUB-SCALE OF THE NASA TLX MAKES TOWARDS OVERALL WORKLOAD BASED ON STANDARDIZED BETA COEFFICIENTS OBTAINED FROM MULTIPLE REGRESSION ANALYSIS.	24
TABLE 3 – MAPPING OF PARTIAL ETA SQUARED AND R_{BIS} ONTO COHEN’S D AND THE CHARACTERIZATION OF EFFECTS AS ‘SMALL’, ‘MEDIUM’ AND ‘LARGE’.	27
TABLE 4 – EFFECT SIZE ANALYSIS OF PERFORMANCE TIME SHOWING THE ‘MEANINGFULNESS’ OF ANY DIFFERENCES EXTANT BETWEEN NOVICES AND EXPERTS.	27
TABLE 5 – EFFECT SIZE ANALYSIS OF PERFORMANCE ACCURACY (SA) SHOWING THE ‘MEANINGFULNESS’ OF ANY DIFFERENCES EXTANT BETWEEN NOVICES AND EXPERTS	28
TABLE 6 – EFFECT SIZE ANALYSIS OF WORKLOAD SHOWING THE ‘MEANINGFULNESS’ OF ANY DIFFERENCES EXTANT BETWEEN NOVICES AND EXPERTS	28

1 Executive Summary

This report completes a suite of papers delivered during 2006 that were concerned with design issues as they related to so-called 'Command Wall systems'. This report is about whether the results gained from novices in previous studies can be generalised to an expert (i.e. military) population.

A considerable corpus of Human Factors research, military and otherwise, is conducted on novices despite their findings being generalised to expert user groups. This report deals with the extent to which novices are 'different' or 'similar' to experts, with particular regard to Command Planning studies.

This study was based on the same Command Planning task(s) used in previous research (HFI DTC 2006, a & b). It uses two groups of participants: The first group was selected from a cohort of Brunel University undergraduates (matching the profile of participants in the previous studies). The second group were serving military personnel based at the British Army's Land Warfare Training Centre in Warminster. A simplified command planning task, based on the Combat Estimate, was undertaken by both groups. Psychological variables such as situational awareness (SA) and mental workload, as well as task performance, were measured.

Whilst experts and novices are clearly not identical there appears to be enough commonality (with some important caveats in place) for them to be used interchangeably in Command Planning studies. The findings show:

- Novices may be somewhat quicker than experts but experts are more accurate.
- Differences in the 'quantity' or 'extent' of probe recall performance and workload are in evidence (experts and novices differ in the relative levels of their measured performance) yet within that the 'pattern' or 'type' of results obtained is broadly concordant (despite differences in the level of performance, the pattern of findings between experts and novices is actually similar), however,
- the critical difference appears to lie in structural and elemental differences in the participant's situation model (which we would expect to differ between experts and novices).

We conclude that novices can be used to good effect in this experimental context but it is important to be alert to critical differences that do exist between the two groups (e.g. Situational Awareness content). We recommend that this issue is revisited periodically should the context to which experts and novice participants are used change.

Although experts are frequently interchanged for novices in military research the potential exists for differences to be so profound that the results are rendered virtually meaningless. An important safeguard against this eventuality (and risk) is to conduct analyses such as this and to continue to do so when warranted.

2 Introduction

2.1 Aims and Objectives

This report presents the results of a follow-on study based on two earlier deliverables. The first of these, entitled “Report on C4i Study: Command Wall System versus Conventional Paper and Radio Based Techniques” (HFI DTC, 2006a) comprised of an experimental study conducted to examine the effects of three C4i techniques; command wall technique (an electronic method embodied by the Brunel command wall system), a basic paper map technique and a more traditional radio and map technique. The task of the participants was to undertake a Battlefield Area Evaluation (BAE) under these three experimental conditions. The aims of this prior study were:

- To discover whether there were significant differences in time, error, mental workload and situational awareness between the three systems.
- To discover any improvements that could be made, from a user perspective, to the Brunel C4i system then under development.

A corollary of the first study was the means by which information contained within the BAE was communicated and subsequently represented, in other words:

- A comparison between whether information is Pushed to the commander or Pulled by them.
- A comparison between whether information remains available for the duration of the study or whether it is only available for a short time.

These two questions formed the basis for a second companion report, entitled “Using an Electronic C4i System to Examine the Effects of Information Source and Decay” (HFI DTC 2006b). Both reports, combined, provided several valuable insights that fed into the design of the prototype Command Wall system hosted in Brunel University’s BIT Lab, as well as highlighting a number of interesting human performance issues.

The aim of this, the final report in the set, is to find out how generaliseable these findings are bearing in mind that that the results were derived from a novice participant pool and are intended to relate to serving military personnel (i.e. experts). The research questions are twofold:

- To discover whether or not there are significant differences between experts and novices and, if so, what these are.
- To discover whether the results derived from novice participants can be generalised to a military population and to what extent.

A large proportion of Human Factors studies are undertaken with novice participants. Paradoxically, a similarly large proportion of Human Factors contexts are those in which

experts, not novices, are the end users. There are several compelling reasons for this situation. The first reason is pragmatic: experts are quite often members of operational staff making it difficult and costly to access them (if access them at all). The second reason is more theoretical: the findings derived from novice participants are quite often generalisable to experts, in other words, the similarities between experts and novices (rather than the differences) are important. The extent of this generalise-ability, or similarity, is, however, dependent upon a whole range of factors unique to the context, to the nature of expertise and to the expert (or novice) themselves. There is no universal answer as to whether findings from all novices are generalisable to 'all' experts (e.g. Dreyfus & Dreyfus, 2005; Arnold, Cooper & Robertson, 2005). The reason for considering this question in a specific domain (military research) and context (command planning scenario) is, as a result, justified.

2.2 Expertise

Expertise can be defined (variously) as:

“having, involving, or displaying special skill or knowledge derived from training or experience” (Merriam-Webster, 2007).

Expertise seems to relate to “the capacity for carrying out complex, well-organised, patterns of behaviour smoothly and adaptively so as to achieve some end state or goal” (Reber, 1995). Some further characteristics include “experts are faster than novices at performing skills; experts perform their tasks almost error free; experts have superior short term memory and long term memory; experts’ problem representation is deeper, more principled, than that of novices, who tend to build superficial representations of a problem.” (Chi & Glaser & Farr, 1988). Such findings are intuitive at a practical level and largely undisputed within the academic literature.

There are a range of theoretical positions adopted in order to explain expertise and there is a considerable degree of conceptual overlap between them. Rasmussen’s (1983) Skill-Rule-Knowledge taxonomy provides a particularly useful descriptive characterisation of expertise, or at least some aspects of it, as explained below.

Skill based behaviour refers to the kind of smooth, well organised patterns of behaviour exhibited by experts. Skill based behaviour has the appearance of being carried out with minimum effort and almost automatically (hence the term ‘automaticity’). A large range of different types of behaviour can be enacted in a skill based manner, from physical skills, such as handling a pen, to more conceptual skills, such as using the Combat Estimate. Skill based behaviour infers that the environmental stimuli, that cue action, are well learned. In fact, they may not even be required at all. Instead, the expert may rely largely on ‘expectancy’, predicting (more often than not correctly) what is going to happen before it actually has and, furthermore, supplying the appropriate corrective, or other response, without requiring ‘knowledge of results’ or feedback. Skill based behaviour relies on a very comprehensive mental model that allows the individual to be closely coupled to the dynamics of the environment. In so doing, skill based behaviour finds analogues in manifold issues related to mental models, working memory (e.g. Ericsson & Kintsch, 1995), practice theory (e.g. Lave, 1996) and a whole range of other

cognitive phenomena. Skill based performance (and the underlying cognitive model) arises as a result of considerable exposure/practice to the task; some researchers suggest that anything up to 10,000 hours may be required (a surprisingly easy target to reach for many day-to-day tasks, such as driving for example; Ericsson, 1996; Charness, Krampe & Mary, 1996; Starkes et al., 1996).

Rule based behaviours are somewhat different. Sometimes the dynamics of the environment but, more often perhaps, the incipient state of the individual's internal model, require cognisance of a range of stimuli and a range of suitable responses. Rule based behaviour, to use a computing metaphor, takes the form of a 'look up table' of stimuli and a similar choice of behaviours. Rule based behaviours conform (approximately) to a kind of IF-THEN process. In between IF and THEN exists what Vicente (1999) refers to as a 'black box', a simplified mapping between stimulus and response that enables the individual to use previously successful behaviours in similar analogous situations.

The sorts of mental processes underpinning skill and rule based behaviours are problematic in the study of expertise:

"If one asks an expert for the rules he or she is using, one will, in effect, force the expert to regress to the level of a beginner and state the rules learned in school. Thus, instead of using rules he or she no longer remembers, as the knowledge engineers suppose, the expert is forced to remember rules he or she no longer uses. ... No amount of rules and facts can capture the knowledge an expert has when he or she has stored experience of the actual outcomes of tens of thousands of situations." (Dreyfus & Dreyfus, 2005).

What this describes are knowledge based behaviours. These are antithetical to the highly detailed mental representation underlying skill based behaviours and the 'black box' of IF-THEN rule based behaviours. Knowledge based behaviours are effortful, conscious, deliberate and often achieve success through trial and error; a real life form of hypothesis testing. Knowledge based behaviours rely on declarative knowledge, which can be consciously accessed, and the retrieval of stored knowledge. Novice behaviour could be characterised as knowledge based. An everyday example of knowledge based behaviour might be the novice driver, who unlike the expert, has to constantly look down at the gear lever to gain feedback as to what gear they are in. Such a characterisation is valid to some extent but does not tell the whole story. Expertise is more than skill-based behaviour.

Remaining with the Cognitive Systems Engineering perspective expertise can be seen as:

- Constructive (it is not the retrieval of pre-planned solutions from memory, rather, it relies on a mutually dependant interaction between expertise and context);
- Tailored (to specific situations; expertise connotes that solutions can be generalised and/or abstracted to similar/diffuse problem areas);
- Action based (rather than a characteristic possessed by an individual).

Expertise is the ability to:

“generate a contextually tailored sequence of cognitive activities that is appropriate for the present situation” (Vicente, 1999). Or, in the words of Vicente (1999), expertise is “like a ‘bag of tricks’, specific enough to be applicable to a particular task, but general enough to be relevant to various situations. As workers become more experienced, they increase the number of bags of tricks in their repertoire” (p. 186). Rasmussen’s (1983) decision ladder illustrates, conceptually, how full the bag of tricks may be, or how expertise may work. It is based on a rejection of the implicit theory underlying prior information processing views (that decision making is linear), in favour of a form of non-linear determinism. Decision making and, with it, expertise, can be seen to be temporally emergent (Vicente, 1999).

Rasmussen’s (1983) decision ladder takes a linear model of decision making and bends it in half. The effect is to emphasise the constructive nature of decision making and the numerous short cuts and shunts that experts use (following the rubric of Skill-Rule-Knowledge based behaviours). This also helps to account for the speed and relative absence of errors seen in expert performance mentioned above (Chi & Glaser & Farr, 1988). Figure 1 illustrates that while novices, couched in knowledge based behaviours, will indeed have to traverse the full height of the decision ladder in order to reach the point of task execution, experts, on the other hand, use all manner of shortcuts. The cognitive representations underlying skill based behaviour represent deeper, more principled problem representation and the repertoire of rules underlying rule-based behaviour are similarly enhanced. These characteristics enable more shortcuts to be taken. There is also an accompanying meta-cognitive element, the expert knows when and how to deploy short cuts and in cases when they cannot be used, their declarative knowledge is likely to be better (a characteristic of expertise is good problem solving).

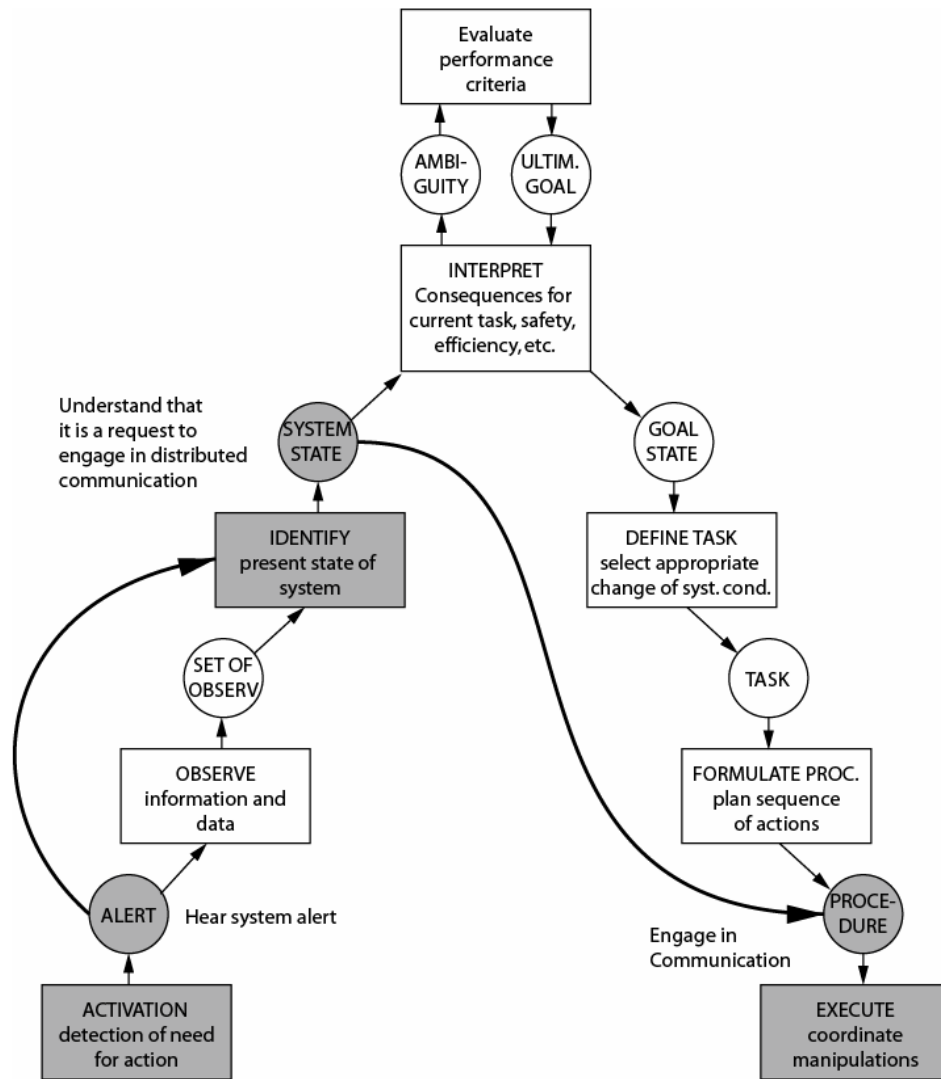


Figure 1 – Rasmussen’s decision ladder. Novice performance would proceed up and down the full height/length of the decision making processes. Expert performance (shown) relies on shortcuts.

2.3 Experts and Novices

The picture painted of expertise might seem to suggest that experts are likely to be radically different from novices. It is certainly the case that while novices will use effortful and time consuming knowledge based behaviours, experts will be able to aggregate more of these behaviours at the skill based level. It is the case that the mental/cognitive representations underpinning the kind of behaviours emitted by experts are considerably different from novices. It also seems likely that novices will find a Command Planning task more effortful (and be less accurate) than military experts. There is a key question to ask: are these differences due to ‘extent’ (i.e. quantitatively different amounts of the same phenomena) or type (i.e. qualitatively different phenomena underpinning performance all together). In other words, do the similarities between experts and novices justify their interchange-ability or do the differences invalidate them?

The literature is fairly equivocal on this point; experts can be substituted for novices but is dependant on the context, the variables of interest and various other caveats. One of the founding papers in Human Factors (Grether, 1949) relates well to the task and context under consideration here. The experiment was concerned with the time/accuracy performance of different avionics displays (for present purposes substitute ‘avionics instruments’ with ‘command planning products’ and the synergy will be apparent). Grether, in the language of 1949, reports that, “College men¹ without [...] experience showed virtually the same pattern of results in this study as highly experienced USAF pilots” (p. 372). There were of course large differences in ‘extent’ but not ‘type’ (in other words, the same design decision could be reached with both cohorts). At least, that is the pattern of results obtained for human performance for relatively simple tasks. The same pattern is not in evidence for more complex aspects of decision making, most notably situational awareness. Randel, Pugh and Reed (1996) found that “novices have not had the experience needed to develop complex models of potential situations, they tend to over-generalise and are unable to readily see how various rules of thumb may conflict” (p. 595). For more straightforward tasks there may not be a great deal of difference, or if there is, it is likely to give experts the ability to ‘do the same things better’. As situations become more complex, however, Randel et al (1996), as well as the wider literature on this topic, suggests that experts can not just do the same things better but ‘better things’ altogether.

The series of Human Factors studies in progress on behalf of the HFI DTC within Brunel University use students as participants and with notable success. It is due to the increasing amount of literature around the area of expertise that we felt the need to complete a suite of papers focused on command planning in which this question of expertise could be considered.

This study will compare the performance of novice participants, on an identical task, to that of serving military personnel (experts). The research questions are twofold:

1. Firstly, are there any significant differences between the two groups of participants and what is the nature of any differences.
2. Secondly, to discover whether or not the results of novice participants can be generalised and applicable to experts drawn from a military population.

The following sections deal with the formal experimental design, the results gained from the study, concluding with the implications of the findings and recommendations stemming from them.

¹ i.e. undergraduates/novices.

3 Methodology

3.1 Design

The study is based around the Combat Estimate planning technique. Participants acting in the role of Commander were required to perform a simplified version of this planning process, which is comprised of seven steps (or questions):

1. What are the enemy doing and why?
2. What have I been told to do and why?
3. What effects do I want to have on the enemy and what direction must I give to develop my plan?
4. Where can I best accomplish each action/ effect?
5. What resources do I need to accomplish each action/ effect?
6. When and where do the actions take place in relation to each other?
7. What control measures do I need to impose?

The between subjects variable has two levels: novices, represented by participants drawn from the student population at Brunel University and experts, drawn from full-time military personnel at the British Army's Land Warfare Centre in Warminster. The independent variable of expertise acted on the following measured dependent variables:

- Task performance (time taken to complete the task);
- Situational awareness (recall accuracy);
- Situational awareness (Critical Decision Method; Klein, Calderwood & MacGregor, 1989);
- Self report mental workload (NASA TLX self-report questionnaire; Hart & Staveland, 1988).

The total time to complete all seven phases of the Combat Estimate task was measured and recorded. The Combat Estimate task enables the participant to organise, structure and represent their Battlefield situational awareness. This representation was compared to an 'ideal' answer provided by a Subject Matter Expert (SME). Thus, in broad terms, actual and ideal SA could be measured. In addition, the Critical Decision Method (CDM) was used in a self-report questionnaire format in order to elicit information on expert decision making and the items of knowledge that comprised the individual's state of SA. The CDM was used to construct knowledge networks. Mental workload is assessed using the NASA TLX self-report questionnaire administered after the experimental task

and anchored to an earlier administration (in the context of an abbreviated practice trial) that occurred at the beginning of the trial. All participants received standardised briefing, debriefing and practice.

3.2 Participants

Twenty participants took part in the study and assumed the role of commander. The mean age of the experts in the study was 37.6 years (minimum of 24 years, maximum of 50 years). The mean age of the novice participants was 23.4 years (minimum 20 years, maximum 26 years). As might be expected, the experts are somewhat older on average than novices (although the broadly comparable minimum ages are suggestive of greater upward spread for the experts). The remit of the study is to compare the performance of these two different populations. Any differences in age are still consistent with this aim and arise as an artefact of 'expert' military populations, just as reduced age is an artefact of the novice participants available to draw upon at Brunel University.

The sample size and between subjects design provide a power level of approximately 0.64 for detecting effect sizes in excess of $d = 0.8$. Thus the study provides a good chance of detecting a large (and meaningful) in excess of $d = 0.8$ effect should such an effect actually exist in the population.

3.3 Materials

3.3.1 Participant Instructions

The task required the participants to plan a mission using the Combat Estimate technique. The mission was described as follows:

"Execute a concentrated and simultaneous operation to disrupt named suspects (Bravos) by searching their houses (Alphas) in order to gather evidence to disrupt and dislocate enemy factions".

This study is concerned with the preparatory and planning phases of this mission, this includes eight stages which are summarised below:

- Q1.1 Conduct a terrain analysis of routes, buildings and sensitive areas within the Battlespace, all of which need to be marked up appropriately.
- Q1.2 Conduct a threat analysis. This involves identifying and highlighting known and probable dicking (enemy surveillance). This stage also involves discovering all of the enemy's previous movements, actions and attacks as well as their current capabilities and contacts.
- Q2. The commander must reiterate and confirm the superior officers' intent, including both specified and implied missions. Constraints on the mission also need to be taken into account.

- Q3. Involves determining mission intent, in other words, working on the ‘battle winning idea’.
- Q4. Involves identifying areas within the battle-space that may need extra support.
- Q5. Involves identifying and allocating the required resources for the mission.
- Q6. Involves developing a course of action.
- Q7. Involves identifying any control measures necessary.

These planning tasks took place on a standard PC laptop (supplemented with additional paper based data collection). Participants had to use graphical tools to annotate the map of the Battlespace in accordance with the tasks above, copies of which were also held on the computer as were databases containing the dependant measures (questionnaires).

The scenario and most of the supporting materials were derived from a handbook designed to accompany Combat Estimate training (Land Warfare Collective Training Group, 2005). The map of the experimental Battlespace is taken directly from this handbook and reproduced in digital format for the purposes of the study (as shown in Figure 2).

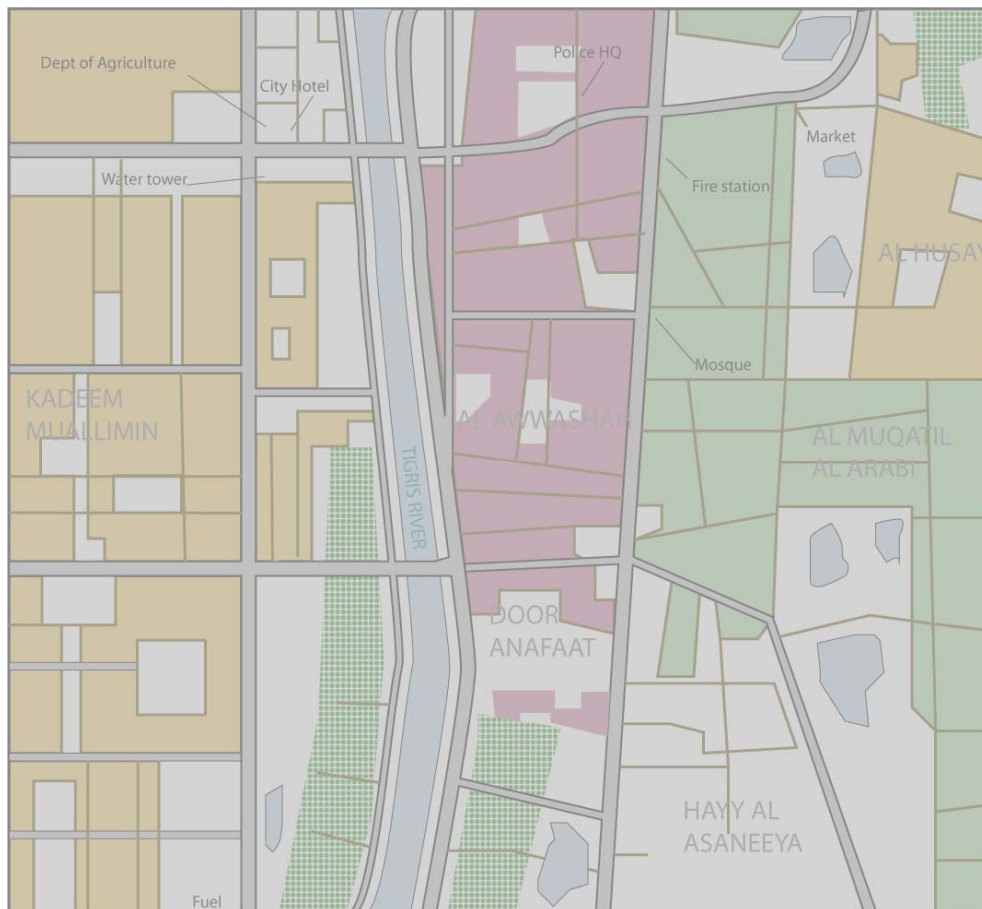


Figure 2 – Representation of the urban battlespace area.

3.3.2 Dependant Measures

The following self report measures were instantiated on a database enabling the participant themselves to enter their responses directly into the PC they were using within the study:

- The NASA TLX (Hart & Staveland, 1988) questionnaire was used to provide a quick measure of self-report mental workload.
- The Critical Decision Method (CDM; Klein, Calderwood & MacGregor, 1989) was employed to provide a structured knowledge elicitation paradigm. The CDM interview protocol was converted into a form of hybrid questionnaire format (the participant typed in their responses to the probes with additional prompting/questioning from the experimenter).

3.4 Procedure

3.4.1 Phase 1 – Pre-Briefing

- Participants were introduced to the aims and objectives of the research and who was sponsoring it. Informed consent was then requested. Ethical, health and safety aspects of participation were dealt with and completed.
- The participant was then given a copy of the commander's instructions to read and any issues and/or questions were dealt with by a member of the study team.
- The participant then undertook a short practice trial to ensure that they were familiar with the experimental paradigm.

3.4.2 Phase 2 – The Combat Estimate

The Commander then begins the first of the Combat Estimate's seven questions:

The participant begins a timer and then answers the first stage of the first question "*what are the enemy doing and why?*" In this first stage the participant marks up a situation overlay, graphically representing the important factors within the Battle-space area. This task was completed using the laptop PC.

The participant begins a timer before beginning to answer the second stage of this question which involves textually representing the important assets within the situation. This question is answered on a blank piece of paper provided labelled Question One. The participant again stops the timer when they have completed the question and records the time. This self timing continues throughout the experiment.

Question two, "*what have I been told to do and why?*", is then answered in a textual form on a blank piece of paper labelled question two.

Question three, “*what effects do I want to have on the enemy and what direction must I give to develop my plan?*”, is then completed on a blank sheet of paper labelled question three.

Question four, “*where can I best accomplish each action/effect?*”, is again split into two stages: the first stage consists of using a computer to annotate a power point slide labelled ‘question four’ with a set of TAI’s (Target Areas of Interest) and NAI’s (Named Areas of Interest). The second stage of this question then involves the participant incorporating this information into a Decision Support Overlay (DSO) matrix which takes the form of a spread sheet.

Question five, “*what resources do I need to accomplish each action?*”, was essentially a multiple choice question; participants were given a range of resources that were available to them and they had to allocate the smallest number of resources possible to the correct locations and functions throughout the Battlespace. This was done textually on a sheet of paper labelled Question Five.

Question six, “*when and where do the actions take place in relation to each other?*”, again requires the annotation of a power point presentation slide on a computer, this time with the participant’s chosen course of action.

Question seven, “*what control measures do I need to impose?*”, is again a two part question. The first part involves annotating a power point slide with any control measures that the participant feels will be needed in order to ensure the mission runs smoothly. The second part consists of textually explaining, on a sheet of paper labelled ‘question seven’, why they have placed the control measures in those specific locations, and what their role is.

A member of the experimental team acts in the role of Technical Facilitator dealing with any procedural or technical issues throughout the trial. The experimenter is also responsible for the archiving and collating of the various power point slides and textual answers provided.

3.4.3 Phase 3 - Questionnaires

After the Combat Estimate task is complete the experimenter administers the two self-report methods (the NASA TLX and CDM questionnaires). Upon completion of these, any further questions and participant payment, the study was concluded.

4 Results

The current study uses 20 participants, 10 novices (Brunel University undergraduates) and 10 experts (serving military personnel), in a between subjects design in order to examine the effect of expertise on the ability to perform a simplified combat estimate task.

4.1 Time

At first glance (*Figure 3*) it appears that the total time taken to complete the task is considerably longer for the experts (Med = 5734 ms) compared to the novices (Med = 3999 ms). Such a finding, at face value, does not appear to sit comfortably with the more skilful and adaptive nature of expert performance described above (e.g. Chi, Glaser & Farr, 1988; Grether, 1949). Clearly, such a supposition needs to be checked.

Further analysis reveals that any differences which are visually manifest in *Figure 3* are not supported statistically. A Mann-Whitney U test² failed to detect a significant difference at the 1, 5 or 10% level for total task completion time between experts (mean rank = 11.8) and novices (mean rank = 9.2): $U = 37$, Exact $p = 0.35$. Of course, this does not mean that there is literally zero influence of expertise on task completion time, it is the case, rather, that a large and meaningful effect size was not detected. *Figure 3* provides additional clues as to the nature of this finding. The 95% confidence interval of the difference between the median scores obtained in the novice condition is 3268ms/6546ms (a range of 3278ms) compared with that of the expert condition, 4523ms/6377ms (a range of 1854). Thus a far greater range of median times are in evidence within the novice condition compared with a much 'tighter' range of values for the experts. Expert performance would be expected to exhibit this property.

² Non parametric techniques are used throughout the study due to the relatively small sample size and a desire to use conservative tests that do not make any underlying assumptions about normality. Note that multiple regression is used in section 4.3 as a means of rank ordering rather than to infer association with underlying assumptions about data normality.

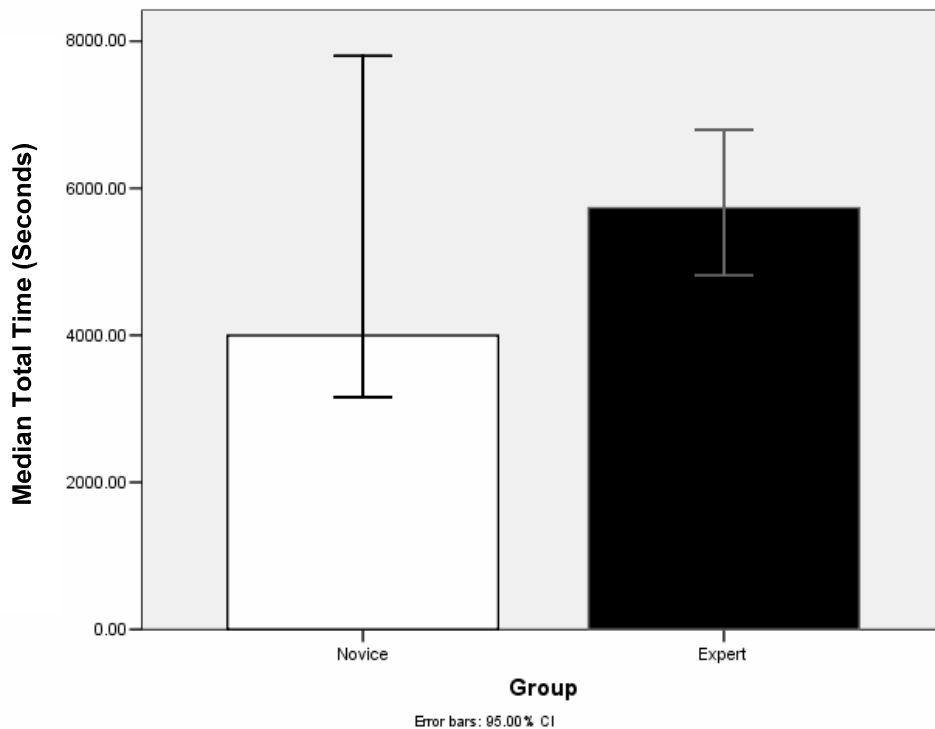


Figure 3 – Bar chart illustrating the median total time taken by the expert and novice groups.

Separating the results into their constituent individual task(s) and subjecting them to a similar statistical treatment reveals no further significant differences in time performance. In all cases the comparison(s) between expert and novices failed to reach significance at either the 1, 5 or 10% level. There was one exception. A significant difference was detected for Question 6 (*"When and where do the actions take place in relation to each other"*). The mean rank for the novice group is 7.7 whereas the expert group is 13.3, a difference that a Mann-Whitney test was able to detect: $U = 22$, Exact $p = 0.04$. For this particular task the novices were significantly quicker.

Question 6 of the Combat Estimate is a relatively challenging task requiring comprehension and prediction on the part of the participants. One might expect expert participants to have access to more knowledge and expertise with which to apply to this task, thereby making it take longer. In addition, there may also be a cumulative effect of task performance in that the outputs of earlier tasks provide more for the expert to contend with in Question 6 compared to the novices (who may not extract the same information or perceive the same cues as experts from the same data). It may, on the other hand, be that novices approach this task in a quantitatively different, perhaps better way than the experts; although this is unlikely given the results that follow.

4.2 Situational Awareness

4.2.1 Accuracy

Participants produced eight outputs corresponding to the seven questions of the combat estimate task (Question 1 had two outputs). All of these outputs were compared against an 'ideal' set of representations (produced by an SME). The concordance between actual and ideal enabled an accuracy measure to be obtained. This measure relates, at a superficial level, to the participants' Situational Awareness (SA). The more concordant the representations supplied by novices and experts were to an 'idealised' version, the better the participant's SA could be said to be. This form of logic underpins the Situational Awareness measure SAGAT (Endsley, 1988). In the present case the accuracy metric is a simple percentage correct score.

When all of the accuracy scores for each of the representations were summed and a median value obtained it was observed that the novice group achieved a median accuracy of 26.25% compared to 43.75% for the experts; the experts were more accurate than the novices. This difference was statistically significant at the 5% level. The mean rank for the experts was 7.5 and the novices 13.5: $U = 20$, Exact $p = 0.02$. Some comfort can be drawn from the fact that regardless of level of expertise all participants were able to achieve some level of accuracy. In other words, the task was certainly not 'beyond' the ability of novices. Having said that, and as one might expect, experts were significantly more accurate on the task at an overall level, suggesting that the extra time taken by the experts yields greater accuracy. This finding is concordant with the wider literature (e.g. Grether, 1949).

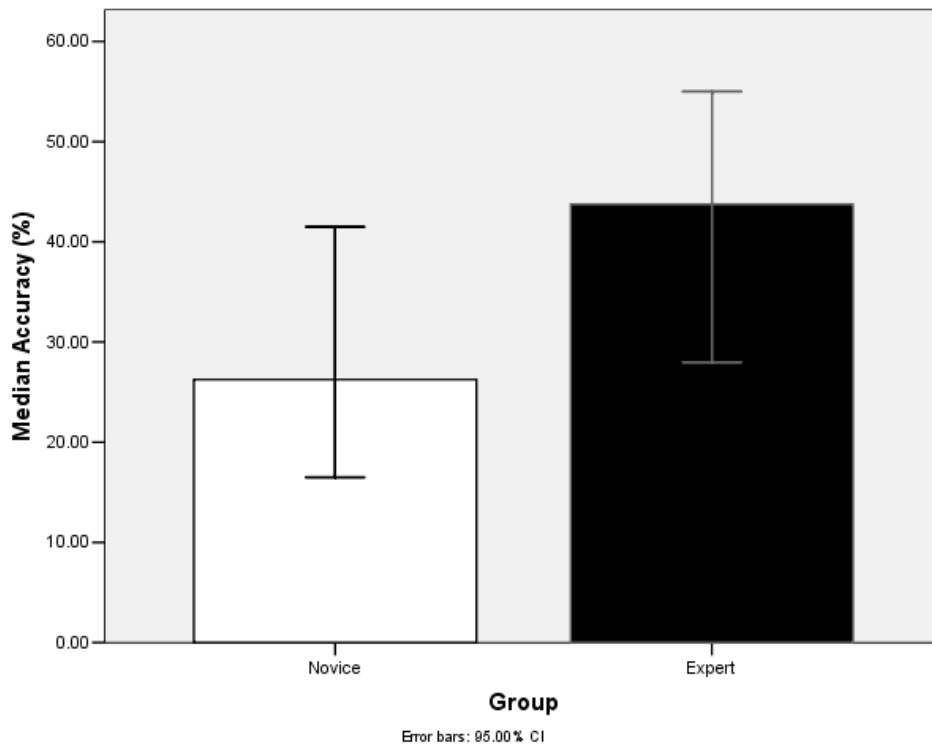


Figure 4 – The accuracy achieved across all seven questions summed to give an overall median accuracy for expert and novice groups.

The level of accuracy achieved by experts and novices was not, however, uniform. Some stages of the combat estimate favoured experts whereas others did not. Significant between subjects effects were detected in task accuracy in:

- Question 2 (mean rank for novices = 7.75, experts = 13.25; $U = 22.5$, Exact $p < 0.05$),
- Question 4 (mean rank for novices = 6.35, experts = 14.65; $U = 8.5$, Exact $p < 0.01$),
- Question 6 (mean rank for novices = 6.6, experts = 14.4; $U = 11$, Exact $p < 0.01$) and
- Question 7 (mean rank for novices = 7.35, experts = 13.65; $U = 18.5$, Exact $p < 0.05$).

In all these cases expert performance was significantly better than novice performance. The same can not be said for:

- Part one of Question One (mean rank for novices = 10.2, experts = 10.8; $U = 47$, Exact $p = ns$),
- part two of Question One (mean rank for novices = 8.5, experts = 12.5; $U = 30$, Exact $p = ns$) or
- Question 3 (mean rank for novices = 9.3, experts = 11.7; $U = 38$, Exact $p = ns$).

This does not mean that there is arithmetically zero difference, as before, it is both the case that the current study failed to detect a large (and, therefore, meaningful) effect and/or one for which the simple effects of random error could not be discounted. This is all the more interesting given that Question One of the Combat Estimate is focused on developing and representing SA of the battlefield (for which there was no significant difference), compared to later parts of the combat estimate which are more focused on course of action development, for which there was. Experts and novices seemed able to perform at comparable levels within tasks that required them to develop SA of the Battlespace but lagged significantly behind experts in terms of putting that SA to use, as Figure 5 shows.

It is apparent that despite differences in the magnitude of difference between the expert and novice participants (e.g. $M = 17.5\%$) the profile of responses (within subjects) is nearly identical. A simple bivariate correlation undertaken using Spearman's Rho shows that expert and novice accuracy rates across the different tasks are significantly concordant: $r_s = 0.88$; $n = 9$; $p = 0.002$. This is reassuring. It is possible to conclude that novice performance, using this simple comparative measure of SA, is broadly comparable to the experts in 'content' if not 'extent'.

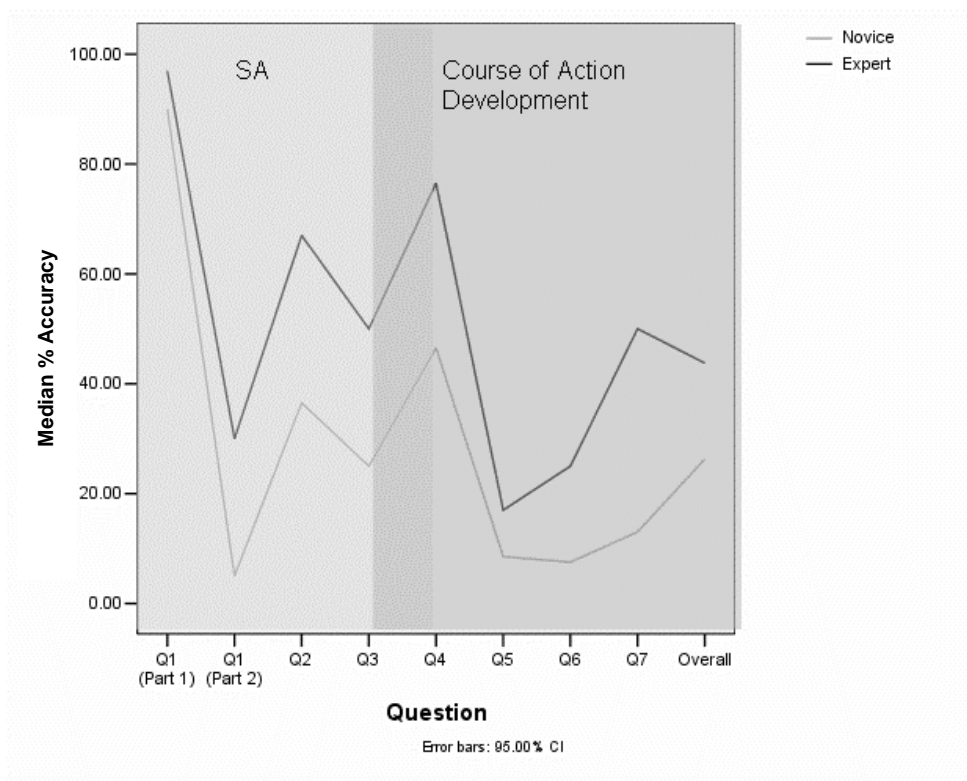


Figure 5 – Line graph illustrating expert and novice median accuracy (%) across the different phases of the combat estimate.

4.2.2 Knowledge Networks

The outputs of the CDM questionnaire formed the basis of a network based analysis. The purpose of this analysis is to probe deeper into the question of SA by creating aggregated situational models for experts and novices. As a modelling exercise this analysis is somewhat different from the inferential approach used elsewhere, yet it still provides a detailed descriptive account of SA content and differences as a complement to the simple accuracy measure described above.

4.2.2.1 Data Reduction and the Creation of the Networks

The output of the CDM semi-structured interview, administered after the experimental task had been completed, is a written transcript. The transcript contains an individual account of the knowledge used and decisions made during the task. Steps have to be taken to move the analysis from the individual/specific level to that of the general/group level. To that end, the transcripts were subject to a five stage process of refinement and data reduction in order that the resultant networks provided a ‘characterisation’ of the group’s SA (expert or novice). The five steps are as follows:

4.2.2.1.1 Stage 1

The participant’s responses to the probe questions were subject to textual analysis. This involved producing a word frequency list comprised of nouns for each participant. Nouns are a robust grammatical category used to describe the “name of a person, place or thing” (Allen, 1984). Nouns are congruent with the idea of ‘situational elements’.

4.2.2.1.2 Stage 2

All words with a frequency of less than five were discarded from the analysis. When plotted on a graph, the word frequency curve approximated (visually) to a form of scree plot. Discarding words with a frequency of less than five removed the situation specific words from the 'tail' or 'scree' of the analysis, as shown in Figure 6. It should be added that this is not a fixed criteria, merely a pragmatic one based on the nature of the research question (requiring a focus on the group rather than individual level) and visual inspection of the data (via the scree plot).

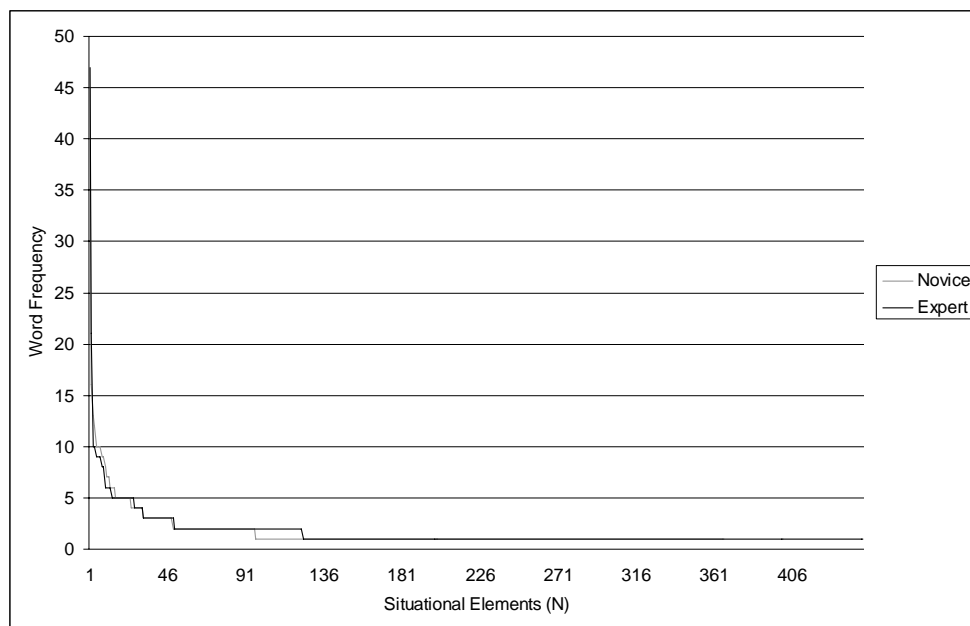


Figure 6 – Word frequency graph illustrating the manner in which the scree (the situation specific words) were removed from the analysis to leave only those elements that occurred consistently.

4.2.2.1.3 Stage 3

The elements (nouns) were checked in order to eliminate repetition (e.g. location vs. locations). Repetitious words were combined.

4.2.2.1.4 Stage 4

The elements were modelled into networks by linking them together using the logical criteria defined below. In this manner, one network was derived for the novice group and another for the expert group. When represented in this way the participants' SA becomes based not only on its 'parts' (the elements themselves) but also its 'interrelations' (the links between elements). Links between elements are established by a set of four logical statements, as follows:

- One element 'has' the property of another element;
- One element 'is' synonymous with another element;
- One element 'requires' the property of another;

- and one element ‘causes’ some property in another.

The transcripts were a form of narrative of important event(s) in the experiment. As such they provided a detailed description of how these situational elements were linked (spatially, temporally, semantically etc.) in terms of the participants understanding of the situation. For example, participant’s literally described how one property affected another (e.g. “yes, engineers were not really a factor” or “I would have to base that on local info and knowledge at the AO³” etc.). The resultant networks that were derived for experts and novices are shown below in Figure 7 and Figure 8. In themselves they are not particularly informative but they do illustrate the extent of interdependence that exists between situational elements and the complexity that this gives rise to.

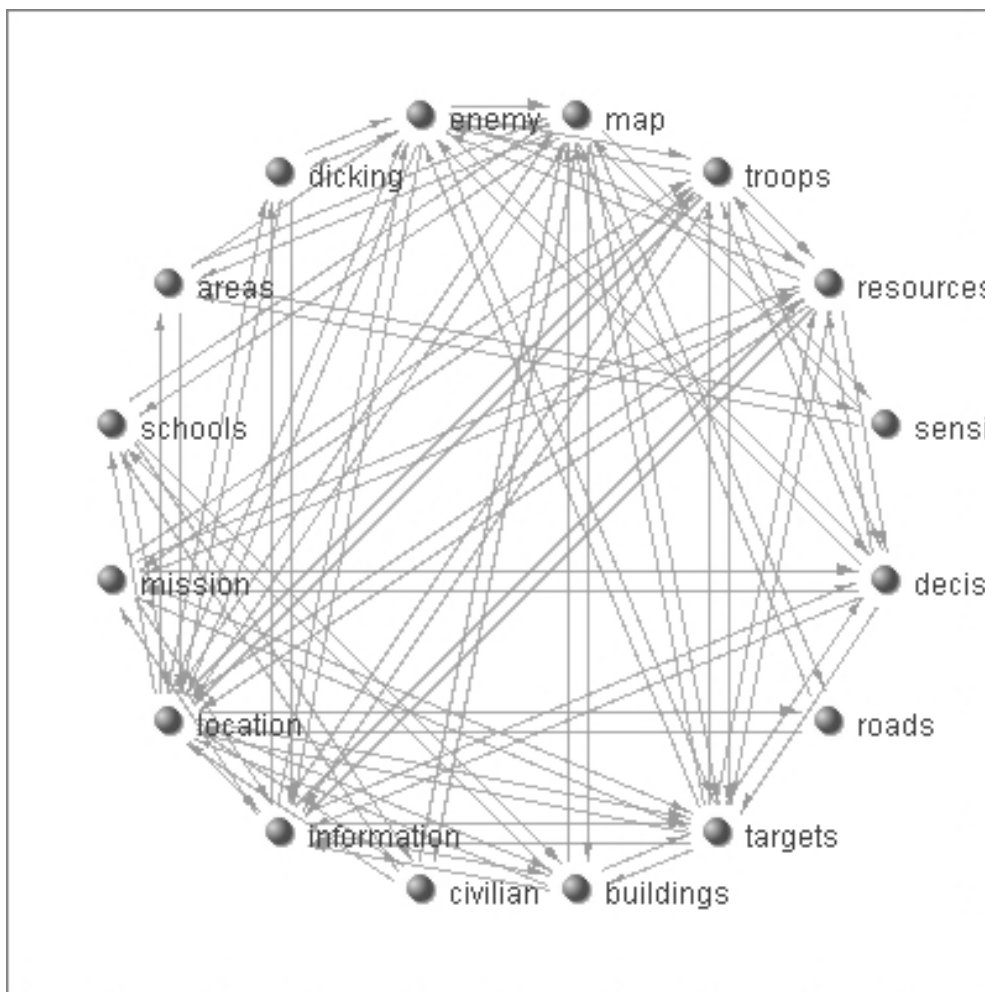


Figure 7 – Knowledge network for the novice group

³ Area of Operations

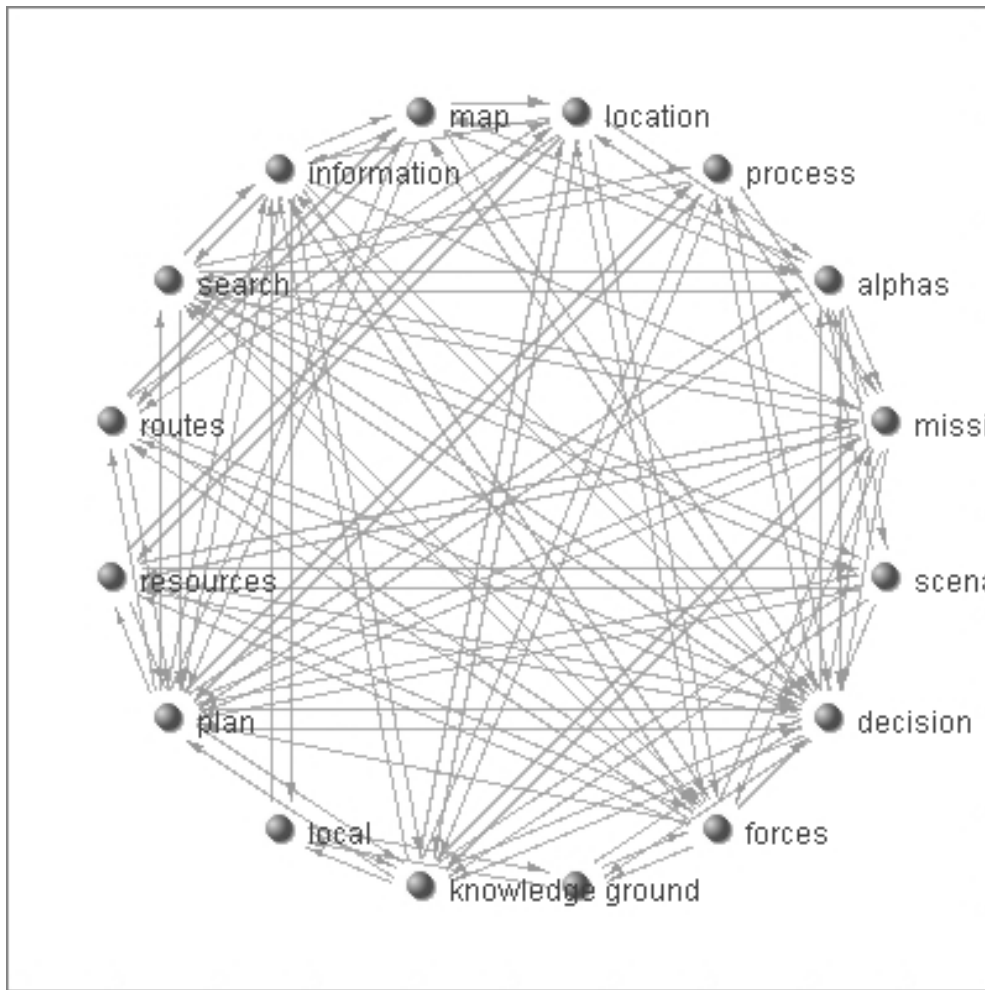


Figure 8 – Knowledge network for the expert group

4.2.2.1.5 Stage 5

The visual complexity of the networks is made tractable through the use of mathematical techniques derived from Graph Theory. This allows emergent properties of the participants' knowledge base, both in terms of 'interrelations' and 'parts', to be distilled and expressed in simple numeric terms.

From within the complexity of Figure 7 and Figure 8 it is discernable that some elements are more heavily interconnected than others and that, overall, the network has a particular level of interconnectivity (i.e. somewhere less than the maximum interconnectivity available, in which every element would be connected to every other one). In other words, the network analysis is not concerned merely with a raw list of elements, instead it takes into account the position that any given element has in the network (its criticality or positional centrality) and the interconnectivity (or density) of the network as a whole.

The first parts of the analysis which use this approach are concerned with the quantity of knowledge and the type of knowledge before moving on to consider the interconnectivity of knowledge in more detail.

4.2.3 Knowledge Quantity

In raw data terms the novices provided a total of 58 sentences comprised of 1343 words into the analysis. The experts provided 56 sentences and 1451 words. This represents an identical median word count of 56 for both groups. The variance around the median value is described by the 95% confidence interval. The lower bound is 47 and 46 for novices and experts respectively. The upper bound is 80 and 86; some experts appear to be supplying more words into the analysis at the upper bound. These differences cannot be regarded as being particularly marked and certainly do not justify an analysis based on trying to detect differences between the means (even if small differences could be detected they do not, based on Figure 9, have any meaningful or practical value).

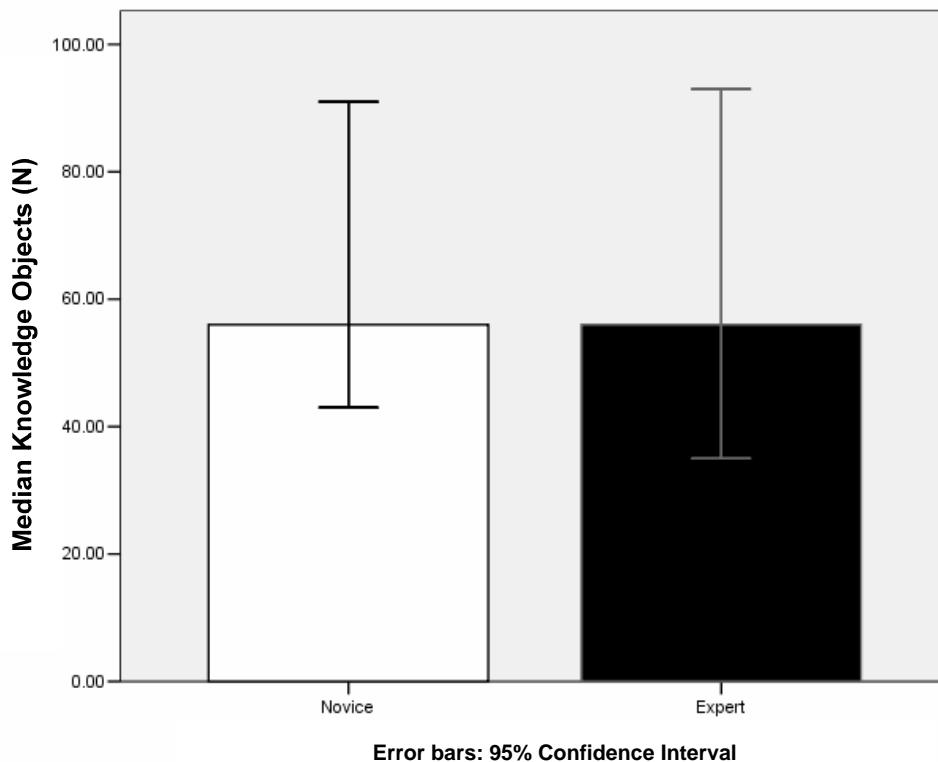


Figure 9 – Bar graph showing the median (and 95% Confidence Interval) for knowledge quantity.

Following step three of the five step process of textual analysis it can be revealed that in terms of the number of situational elements, novices and experts alike provide 16 situational elements for inclusion in the networks above. The lack of difference in quantity, therefore, persists even at this level of analysis. Thus experts and novices, in this experimental scenario, extract what can be considered, for all practical purposes, an identical amount of situational knowledge from the scenarios. The innate ability of novices to extract a given informational content for a scenario such as this appears to be congruent with the experts; the experts, as subsequent sections show, seem to be able to deploy this innate ability in a more targeted and focused manner (Randel et al., 1996). Although the quantity is the same, the type of situational elements extracted differs.

4.2.4 Knowledge Type

Table 1 presents the situational elements derived from this analysis along with their attendant positional centrality⁴ scores (the higher the score the more interconnected and, presumably, important it is).

Table 1 – Comparison between experts and novices in terms of the type of situational elements extracted from the scenario shown with their attendant positional centrality scores (in rank order: most central elements first).

Novice Situational Elements	B-L Centrality	Expert Situational Elements	B-L Centrality
Location	11.24	Decision	10.45
Map	9.91	Plan	10.16
Targets	9.91	Mission	8.74
Enemy	9.67	Forces	8.55
Troops	8.67	Knowledge	8.55
Resources	8.67	Search	8.55
Information	8.67	Information	8.55
Buildings	8.32	Location	8.00
Schools	7.43	Map	7.83
Dicking	7.43	Alphas	7.83
Decision	7.30	Process	7.52
Areas	7.30	Scenario	7.52
Civilian	7.05	Resources	7.52
Mission	6.93	Routes	7.23
Roads	6.82	Ground	6.83
Sensitive	6.12	Local	6.26

The first six (highest ranking) elements are completely different for experts and novices. The novices seem to focus more on external artefacts of the situation, such as location, map, targets, enemy and so on. The experts seem less concerned about artefacts and more focused on ‘concepts’, such as decision, plan, mission, knowledge and so forth. In Endsley’s terms (1995) the novices seem to be couched more at a ‘perception of elements in the environment’ level (i.e. Level 1 SA) whereas experts seem to be concerned principally with ‘comprehending what those elements mean’ (i.e. Level 2 SA) and ‘projecting this understanding in order to anticipate future states’ (i.e. Level 3 SA). This finding is highly congruent with expert performance. Classic studies into expertise (e.g. Chase & Simon, 1973) show that whereas novices are focused on parts, experts are focused on patterns and larger concepts and/or chunks.

Table 1 also shows that only 6/16 (or 38%) of situational elements are shared between experts and novices. This is a clear indication that expert SA is considerably different in content (rather than quantity) than novice SA. Of those elements that are shared, three (location, map and resources) are rated higher for the novices than the experts. One (information) has identical positional centrality and two (decision and mission) are ranked higher by the experts. It is also apparent that the terminology used by experts is more expressive and chunked, which is an additional factor in the apparent lack of overlap. For example, whereas novices speak of ‘roads’ experts speak more generally of ‘routes’, whereas experts speak of ‘alphas’ novices use the more naïve term ‘enemy’.

⁴ The measure of positional centrality used for this analysis is the Bevelas-Leavitt statistic.

These nomenclatural differences are still indicative of a qualitatively different type of SA; the novices being more specific and atomistic, the experts being conceptual and aggregated.

4.2.5 Interconnectivity of Knowledge

The final aspect of expert versus novice SA to be considered relates now to the interconnections between elements rather than the elements themselves. This is an important, arguably neglected part of SA theory. The interconnections between situational elements determine the network's function and bestow upon it the property of emergence (we contend that SA, along with many cognitive phenomena typically viewed in an atomistic, single channel way, are in fact emergent and non-linear).

Network density is a numeric metric based on the comparison between the number of interconnections that are possible (i.e. every node connected to every other node) and those that are actually extant⁵. The results of this analysis show that the novice network is comprised of 16 nodes and 94 interconnections. The expert network is comprised of 16 nodes as well but has 111 interconnections. Thus the density of the novice network is 0.39 compared to the expert network which is 0.46; the expert network is more interconnected, it is more dense.

This finding is again concordant with the wider literature on expertise (e.g. Randel et al., 1996). A greater degree of interconnectivity between situational elements means that they can relate to each other in more ways. The combination of quantitatively different types of knowledge and greater network density gives expert SA completely different (non-linear) properties compared to the novices. This finding would be expected. For experimental purposes it appears that the 'quantity of SA' (to use a crude phrase) can be studied using novice participants. Questions dealing with the 'content of SA' would profit from using expert participants.

4.3 Mental Workload

Mental workload was measured using the NASA TLX self-report questionnaire. The questionnaire was applied twice, once to an anchor task and, again, after the experimental tasks were complete. The results are based on subtracting the second application of the questionnaire from the first. Under this procedure each participant serves as their own workload baseline (thus the results are not confounded by different workload 'starting points' which might be unique to each participant). With this controlled for, the analysis proceeds first at an overall level with each of the six NASA TLX subscales aggregated into a mean workload score.

⁵ Network density is not, however, a simple proportion. Other mathematical characteristics of the network are taken into consideration.

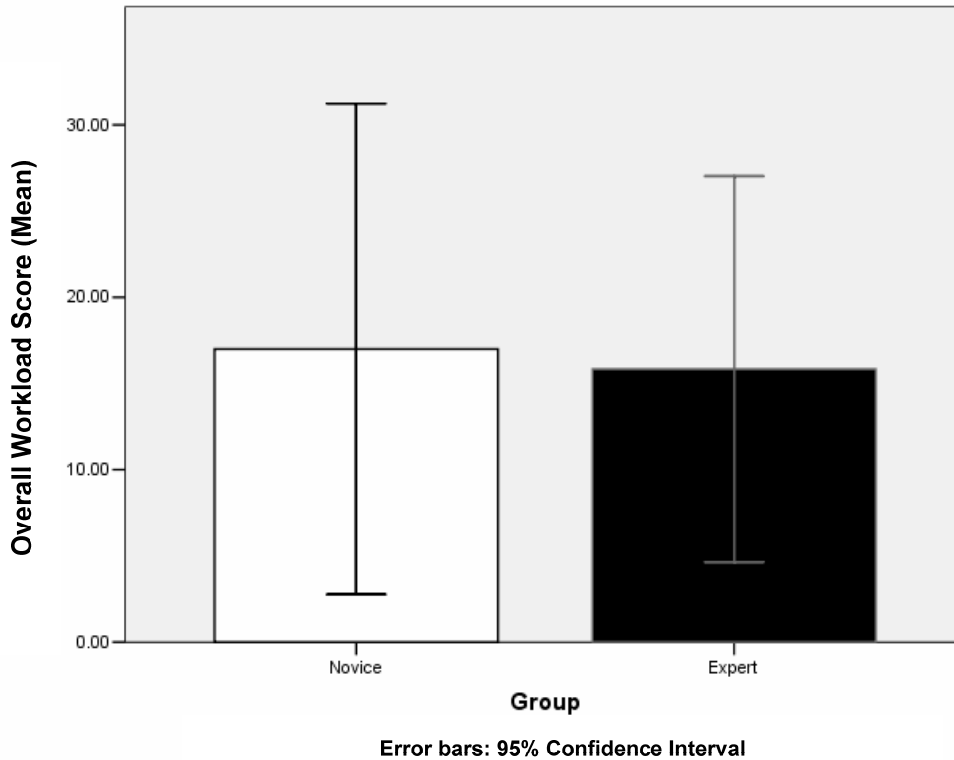


Figure 10 – Bar chart showing overall workload for the novice and expert groups.

Figure 10 shows that there is little practical difference between the mean overall workload achieved by the novice group ($M = 17$; $SD = 19.9$) compared to the expert group ($M = 15.83$; $SD = 15.64$). This difference was not indicative of a significant effect at either the 1, 5 or 10% level: $U = 47$, Exact $p = 0.85$.

The next phase of the analysis considers the relative contribution that each sub-scale makes towards the overall workload score (and how this might differ between the two groups). Multiple regression analysis is used to compute this using a procedure advocated by Warm, Dember & Hancock (1996). The calculation of Beta coefficients within the multiple regression procedure here provides a means to rank order the contribution that each sub-scale makes towards overall workload. The results are shown in Table 2.

Table 2 – Relative contribution(s) that each sub-scale of the NASA TLX makes towards overall workload based on standardized beta coefficients obtained from multiple regression analysis.

Novice (Rank Order)	Expert (Rank Order)
Physical Demand	Performance
Temporal Demand	Effort
Frustration	Mental Demand
Effort	Frustration
Performance	Physical Demand
Mental Demand	Temporal Demand

Although there is little in the way of difference between experts and novices in terms of overall workload, the contribution that each NASA TLX sub-scale makes towards that overall measure does differ considerably. In fact, no one individual sub-scale makes the

same relative contribution within the two groups. For the experts, the top three contributors to overall workload are Performance, Effort and Mental Demand. For the novices this is Physical Demand, Temporal Demand and Frustration. Findings of this nature would be expected. It appears that the novice group's workload is affected by artefacts of the experiment such as time pressure and frustration whereas, on the other hand, experts are more affected by 'task based' factors such as performance, effort and mental demand.

The level of workload experienced in these and the overall workload categories does not appear to be high enough to drastically affect performance nor are any of the differences statistically significant when pairwise comparisons are performed across all the individual sub-scales. The results of this analysis (using a Mann-Whitney U test) and the previous (multiple regression) are rendered in Figure 11. Despite the differences already noted in terms of physical and temporal demand, what is equally striking is the extent of commonality. Indeed, a correlational analysis undertaken using Spearman's Rho was significant, albeit at the 7% level: $r_s = 0.72$; $n = 7$; $p = 0.068$.

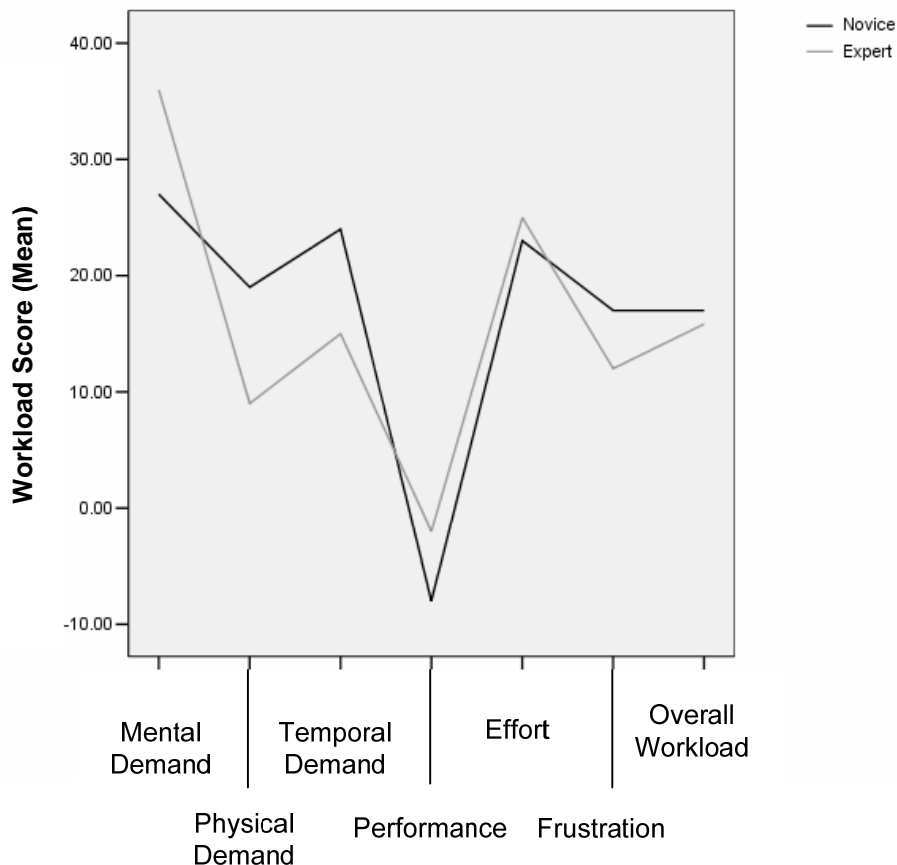


Figure 11 – Line graph showing the results obtained when the anchor and treatment applications of the NASA TLX questionnaire were subtracted from each other for both expert and novice participants.

Taken as a whole, there appears to be some level of reassurance that despite changes in the relative contribution that each sub-scale makes towards overall workload, there are no statistically significant differences between experts and novices. On the contrary, there is

correlational evidence of a degree of statistical concordance between the two groups. For experimental purposes it appears that the 'amount of workload' can be usefully studied using novice participants. Questions dealing with the 'contribution that each workload subscale makes' would profit from using expert participants. Thus the findings present an identical picture to those described in the SA section above.

4.4 Proving the Null Hypothesis

This study is focused on the question of whether novices can be used in experimental studies that relate to experts. This places the findings in an unusual conceptual bind. Statistical procedures, and the underlying hypothetical-deductive approach, are premised on the detection of *differences* between expert and novices, that is, the extent to which any difference could have occurred due to random error (and by implication the probability that the null hypothesis can be rejected).

An equally desirable outcome for this study is that there are no (meaningful) differences between experts and novices, thus the null hypothesis has to be 'proved'. This issue has to be dealt with carefully and the purpose of this section is to explain the conceptual approach used.

Firstly, there is a distinction to be made between the null hypothesis and what Cohen (e.g. 1994) refers to as the 'nil' hypotheses. The critical point is that failure to find support for an experimental hypotheses does not connote that there is precisely nil difference between the variables. On the contrary, there are always differences, however small, between dependent (the variable being measured) and independent variables (the one being manipulated), the question is simply to how many decimal places (Cohen, 1994). Significant differences can always be found given a large enough sample, one large enough to detect effect sizes to the *n*th decimal place. The conceptual response to this, and to 'proving' the null hypothesis, is to consider the effect size along with statistical significance.

An 'effect' is a descriptive term used to describe the amount of difference caused in the dependent variable as a result of manipulating the independent variable. Thus, given a large enough sample, a very small and, for all practical purposes, meaningless effect can be detected. The current study was designed with this issue in mind; the sample size provides an opportunity for large and meaningful effects to be detected if such large and meaningful effects exist in the population as a whole. We assume that a large effect is also an important and meaningful one for practical purposes whereas a small effect is not (thus does not require the effort and expense of recruiting the very large sample size required to detect it). An effect size analysis, therefore, considers not just the presence/absence of a statistically significant finding but also its size and, by implication, its importance. A descriptive example of small, medium and large effect sizes has been proposed by Cohen (1962) as follows:

A small effect size might be the kind of difference in height between 15 and 16 year olds (there would be considerable overlap and it would be difficult to discern, on the basis of height, who was 15 and who was 16). Detecting such an effect size might be of questionable importance for any practical purpose given that, in this case, age is such a poor predictor of height. A medium effect is often described as being discernable by the naked eye, such as the difference in height of 15 and 18 year olds (it would be relatively easy to see, based on height, who belonged to which group, thus age, in this instance, might have some practical benefit as regards predicting height). A large effect has been

described as ‘grossly perceptible’, of a magnitude similar to the difference in height, for example, of 13 and 18 year olds (in this case age becomes an obvious and highly apparent predictor of height). This study is couched at the latter level of ‘confirmatory’ analysis.

Cohen’s d statistic is used as the effect size metric (and is based on the standardized mean difference between expert and novices). This measure can be mapped onto the characterisation of small, medium and large effects as shown in Table 3, Table 4, Table 5 and Table 6. They go on to present an effect size summary, enabling the null hypothesis to be if not ‘proven’ then at least substantiated. Note that the network based analysis is excluded from this treatment as this is a form of ‘modelling’ rather than statistical inference.

Table 3 – Mapping of partial eta squared and R_{bis} onto Cohen’s d and the characterization of effects as ‘small’, ‘medium’ and ‘large’.

Cohen’s d statistic	Effect Size Heuristic
0.2	Small
0.3	
0.4	
0.5	Medium
0.6	
0.7	
0.8	Large
0.9	
1.0	
1.1	
1.2	
1.3	
1.4	
<1.5	

Table 4 – Effect size analysis of performance time showing the ‘meaningfulness’ of any differences extant between novices and experts.

Time	Cohen’s d statistic	Effect Size Heuristic
Q1 (Slide)	0.52	Medium
Q1 (Paper)	-0.13	Small
Q2	-0.42	Medium
Q3	0.51	Medium
Q4 (Slide)	0.28	Small
Q4 (DSO)	-0.50	Medium
Q5	-0.65	Medium
Q6	-1.27	Large**
Q7 (Slide)	0.23	Small
Q7 (Paper)	0.01	Small
Total	0.02	Small

Table 5 – Effect size analysis of performance accuracy (SA) showing the ‘meaningfulness’ of any differences extant between novices and experts

Accuracy (SA)	Cohen's d statistic	Effect Size Heuristic
Q1 (Slide)	0.02	Small
Q1 (Paper)	-0.69	Medium
Q2	-1.09	Large**
Q3	-0.45	Small
Q4	-2.01	Large***
Q5	0.18	Small
Q6	-1.59	Large***
Q7	-1.41	Large**
Mean	-1.67	Large***

Table 6 – Effect size analysis of workload showing the ‘meaningfulness’ of any differences extant between novices and experts

Workload	Cohen's d statistic	Effect Size Heuristic
Mental Demand	-0.4	Small
Physical Demand	0.33	Small
Temporal Demand	0.32	Small
Performance	-0.18	Small
Effort	-0.07	Small
Frustration	0.17	Small
Overall	0.07	Small

* = Statistically significant at the 10% level

** = Statistically significant at the 5% level

***=statistically significant at the 1% level

Note: A large effect size is not automatically statistically significant as it depends on, amongst other things, variance in both sets of control/intervention data and so forth.

Novices appeared to differ to a meaningful extent from Experts in terms of speed and accuracy. Large effects were detected pursuant of the following:

1. Novices were significantly (and meaningfully) quicker when completing Question 6 of the Combat Estimate but
2. Experts were significantly (and meaningfully) more accurate in the task overall and specifically within Questions 2, 4, 6 and 7.

In all other respects there appears to be little in the way of evidence (statistical and effect size) to suggest that experts and novices differ to ‘any meaningful extent’. Whilst it is not possible to conclude that there is zero difference (the so-called ‘nil’ hypothesis) the effect sizes obtained are small, lending credence to the position that even if a larger sample were deployed (and a significant difference detected) it would be relatively meaningless in real-world terms.

5 Conclusions

Whilst it is not generally possible to conclude that in all circumstances novices can be substituted for experts in Human Factors experiments, the balance of evidence in this study suggests that in this specific experimental context their use is justified.

There can be no doubt that experts are superior and, indeed, qualitatively and quantitatively different in some respects to novices. In the current study the application of expertise led to longer task completion times than the novices but also far superior task accuracy. Experts also had a considerably different situational model to that of novices. Although the same amount of knowledge was extant in the network, the type of knowledge was different and the interconnections between the knowledge were greater. Having said that, in this, and most other instances, the differences were ones of extent rather than type. In accuracy and Mental Workload, for example, the novices exhibited statistically similar patterns of response albeit at a reduced level. What seems to be most striking within these results are the similarities rather than differences. The key findings/recommendations that apply to expert versus novice performance in Command Planning tasks are as follows:

- Contrary to the literature, expert and novice performance may differ in task time, with novice performance actually being quicker. This is argued to arise as an artefact of expertise simply taking longer to apply in this type of task.
- Experts, concordant with the literature, are more accurate than novices. Novices are still, however, able to function within the Command Planning paradigm, in other words, the task/context is not 'beyond them' provided that suitable training and practice is provided.
- Although accuracy is greater for experts compared to novices, the pattern of accuracy across different Command Planning tasks is uniform between the two groups. The differences are ones of extent rather than type.
- The structure and type of SA developed by experts and novices is clearly different (and relates to the accuracy findings above). For experimental purposes it appears that the 'quantity of SA' can be studied using novice participants whereas questions dealing with the 'content of SA' would profit from using expert participants.
- The 'amount of workload' did not differ between experts and novices and thus can be usefully studied using novice participants in this context. Questions dealing with the 'type' of workload, in terms of the contribution that each workload subscale makes to the overall workload score, would profit from using expert participants as this does differ considerably. Presumably this arises as a result of the different situational models developed and used during the task but remains a topic for further research.

Whilst experts and novices are clearly not identical there appears to be enough commonality (with some important caveats in place) for them to be used interchangeably in Command Planning studies. The findings help to identify what factors are similar and what are different and appear to be concordant with the wider literature as well (e.g.

Grether, 1949 and Randel et al., 1996). It can be concluded that novices can be used to good effect in this experimental context but it is important to:

- be alert to critical differences that do exist (e.g. SA content) and
- to periodically revisit this issue should the experimental context change.

6 References and Bibliography

- Allen, R. E. (1984). The Pocket Oxford Dictionary of Current English. Oxford: Clarendon.
- Arnold, J., Cooper, C. L. & Robertson, I. T. (1995). Work psychology: Understanding human behaviour in the workplace. London: Pitman.
- Charness, N., R. Th. Krampe, and U. Mayr, 1996, 'The role of practice and coaching in entrepreneurial skill domains: An international comparison of life-span chess skill acquisition.' In, The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games, K. A. Ericsson, ed. Mahwah, NJ: Erlbaum, pp. 51-80.
- Chase, W. G. & Simon, H. A.(1973). Perception in chess. Cognitive Psychology, 4, 55-81.
- Chi, M. T. H., Glaser, R. & Farr, M. J. (1988). The Nature of Expertise. New Jersey: Lawrence Erlbaum Associate Publishers.
- Cohen, J. (1994). The earth is round ($P < 0.05$). American Psychologist, 49, 997-1003.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.
- Dreyfus, H & Dreyfus S. (2005). Expertise in real world contexts. Organisation Studies, 26(5). 779-792.
- Driskell & Mullen (2005). Social Network Analysis. In N. A. Stanton et al. (Eds.), Handbook of Human Factors and Ergonomics Methods (pp. 58.1-58.6). London: CRC.
- Endsley, M. R. (1988), Situation awareness global assessment technique (SAGAT). Proceedings of the National Aerospace and Electronics Conference (NAECON). (New York: IEEE), 789-795.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37, (1), 32-64.
- Ericson, K. (1996). The Road to Excellence - The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games. New Jersey: Lawrence Erlbaum Associate Publishers.
- Ericsson K.A. & Kintsch W. (1995). Long term working memory. Psychological Review. 102, 211-245.
- Gobet, F. (1998). Expert memory: a comparison of four theories. Cognition, 66, 115-152.
- Grether, W. F. (1949). Instrument reading. I. The design of long-scale indicators for speed and accuracy of quantitative readings. Journal of Applied Psychology, 33, 363 – 372.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human Mental Workload (pp. 138-183). Amsterdam: North-Holland.
- Hutchins, E. (1995). How a cockpit remembers its speeds. Cognitive Science. 19, 265-288.

- Hutchins, E. (1995). Cognition in the wild. Cambridge, MA: MIT Press.
- HFI DTC (2006a). Report on C4i Study: Command Wall System versus Conventional Paper and Radio Based Techniques. Yeovil: Aerosystems.
- HFI DTC (2006b). Using an Electronic C4i system to examine the effects of Information Source and Decay.. Yeovil: Aerosystems.
- Klein, G. & Armstrong, A. A. (2005). Critical decision method. In N. A. Stanton et al. (Eds.), Handbook of Human Factors and Ergonomics Methods (pp. 35.1-35.8). London: CRC.
- Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. IEEE Transactions on Systems, Man, and Cybernetics, 19(3), 462-472.
- Land Warfare Collective Training Group (2005). The BG PSO Combat Estimate. Warminster: MoD.
- Lave, J. (1996). The practice of learning (pp. 3-32). In S. Chailkin & J. Lave (Eds.). Understanding Practice: perspectives on activity and context. NY: Cambridge University Press.
- Merriam-Webster (2007). Merriam-Webster OnLine. Available at: www.merriam-webster.com.
- MoD (Aug 2/3rd 2005a). Presentation "Command and Staff Trainer (South) 1 WFR MiniCAST". Land Warfare Centre, Warminster.
- MoD (Aug 3rd 2005b). Presentation "Wargaming: Mastering Your Enemy" (based on the Army Field Manual Vol I (Combined Arms Operations) Part 2 (Jul 98) and 3(UK) Div Wargaming Aide Memoir)". Land Warfare Centre, Warminster.
- Randel, J. M., Pugh, L. H., & Reed, S. K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. International Journal of Human-Computer Studies, 45, 579-597.
- Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 257-266.
- Reber, A. S. (1995). The Penguin Dictionary of Psychology. London: Penguin.
- Stanton, N. A., Salmon, P., Walker, G. H, Baber, C. & Jenkins, D. (2005). Human Factors Methods: A Practical Guide for Engineering and Design. Aldershot: Ashgate
- Starkes, J. L., J. Deakin, F. Allard, N. J. Hodges, and A. Hayes, 1996, 'Deliberate practice in sports: What is it anyway?' In, The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games, K. A. Ericsson, ed. Mahwah, NJ: Erlbaum, pp. 81-106
- Vicente, K. (1999). Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. Mahwah, NJ: Lawrence Erlbaum Associates.
- Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman & M. Mouloua (Eds.), Automation and human performance: Theory and applications. (pp. 183-200). Mahwah, NJ: Lawrence Erlbaum Associates.

