Implementing the Endeavor Space Dimensions

Towards an understanding of perceived complexity in C2 operations

Oscar Bjurling
Jacob Weilandt

Tutor: Björn J E Johansson
Examiner: Arne Jönsson
Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/her own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Oscar Bjurling & Jacob Weilandt
Abstract

The challenge of operating and managing complex and dynamic environments, known as complex endeavors, has become a central issue in the C2 research community. NATO research groups have studied how to combat the negative effects of endeavor complexity on performance. Essential to these efforts is the study of C2 Agility, which is the ability of an entity to cope with change and employ different C2 approaches based on the requirements imposed by—and changes in—the current operational environment. An important aspect in accomplishing this research goal is to study how operational environments are constituted, as this would enable research into how the effectiveness of different C2 approaches is affected by different endeavors. The Endeavor Space model, which represents endeavor complexity in three dimensions, was developed for this purpose. In an effort to continue research on the Endeavor Space, the current study set out to implement the dimensions in a C2 research platform called ELICIT. Three ELICIT scenarios were created to represent different regions of the Endeavor Space. Additionally, the study designed, developed, and tested a prototype self-assessment instrument—the ESSAI—to capture how the Endeavor Space dimensions—Tractability, Dynamics, and Dependencies—were experienced by operators. Eight teams completed the scenarios and rated their complexity using the ESSAI. No significant differences in perceived complexity could be found between the scenarios. However, all Endeavor Space dimensions indicated correlational relationships with perceived difficulty, and most of them correlated with ELICIT performance. This is indicative of underlying patterns that were not thoroughly revealed in the current study. Implications and improvements for future research are discussed.

Key words: Command and Control, C2 Agility, ELICIT, Complexity, NATO, Endeavor Space.
Acknowledgements

We would like to thank our supervisor Dr. Björn J E Johansson and Dr. David S Alberts for their continuous guidance, support, and constructive feedback during the course of this project. We would also like to thank everyone at the Swedish Defense Research Agency (FOI) for their hospitality and feedback, and the conversations we’ve enjoyed. Also, we would like to thank each and every one of our participants, without whom this project would not have been possible. We would also like to extend our thanks to Dr. Peter Berggren and Dr. Erik Prytz from Linköping University, and Per-Anders Oskarsson from FOI, for their statistical advice and guidance. Lastly, we thank our families and friends for their unconditional love and support over the years and throughout this project.

Oscar Bjurling and Jacob Weilandt
Linköping, June 2019
# Table of Contents

1 Introduction  
   1.1 Purpose  
   1.2 Research questions  
   1.3 Hypotheses  
   1.4 Delimitations  

2 Theory  
   2.1 Command and Control (C2)  
      2.1.1 C2 Approach Space  
   2.2 C2 Agility  
      2.2.1 C2 Endeavour Space  
   2.3 Simulations  
      2.3.1 Concepts and Definitions  
      2.3.2 Simulation Typology  
      2.3.3 ELICIT  
   2.4 Team Cognition and Performance  
   2.5 Background synopsis  

3 Method  
   3.1 The Endeavor Space Subjective Assessment Instrument  
      3.1.1 Design rationale  
      3.1.2 Structure and scoring rationale  
      3.1.3 Testing and iteration  
         3.1.3.1 Results and revisions  
      3.1.4 Validation  
         3.1.4.1 Pandemic  
         3.1.4.2 Participants  
         3.1.4.3 Ethics  
         3.1.4.4 Scenarios  
      3.1.5 Design  
      3.1.6 Materials  
      3.1.7 Procedure  
      3.1.8 Results  
   3.2 ELICIT Experiment  

vii
3.2.1 Participants 53
3.2.2 Ethics 54
3.2.3 Design 54
3.2.4 Instruments 54
3.2.5 Equipment and materials 55
3.2.6 Scenarios 56
3.2.7 Scoring 59
  3.2.7.1 ESSAI rating scores 59
3.2.8 Procedure 61
3.2.9 Analysis 62
4 Results 63
  4.1 ESSAI internal consistency results 63
  4.2 ELICIT experiment results 68
5 Discussion 75
  5.1 Method 75
  5.2 Results 80
6 Conclusions 87
  6.1 Future research 88
7 Bibliography 91
Appendix A – Endeavor Space Subjective Assessment Instrument (ESSAI) 97
Appendix B – Pandemic consent form 100
Appendix C – ELICIT trials consent form 101
Appendix D – NASA Task Load Index (TLX) 102
Appendix E – Situation Awareness Rating Technique (SART) 103
1 Introduction

Moving into the Information Age at the start of the 21st century has seen a rapid development of information technology, a drastic increase of available information, and an increase in information channels. This, coupled with the fact that connectedness is a key philosophy of the age, has led to the creation of a complex web of intermingled actors and systems (Atkinson & Moffat, 2005). Therefore, for all its benefits, the Information Age also offers a new type of challenge; the challenge of operating and managing complex and dynamic environments - a challenge previously unobserved in the linear and (comparatively) static environments of the Industrial Age (NATO SAS-085, 2013).

The challenges inferred by complex operations, as defined in the Information Age, have been particularly acknowledged within the domain of Command and Control (C2). The function of C2 is to establish control over a situation in order to safely and successfully direct operations towards a premeditated end state. This has previously been accomplished, with great success, through deconstructing problems into manageable parts and allowing specialized units to manage each subproblem. The historical rationale of the C2 organizations has been to deconflict their organizations to the furthest extent possible to ensure that each specialized unit is only engaged in doing the task relevant to them, thus focusing their efforts to decrease task completion times and achieve more immediate success (Alberts & Hayes, 2003). Although the challenges of the Industrial Age were not trivial in any way, the deconflicted approach was used successfully to manage complicated and considerable challenges. The operational environments facing C2 organizations in the 21st century is, however, an altogether different matter.

The term “complex endeavors” has become central to C2 research and application. Alberts, Huber, and Moffat (2010, p. 16) provide the following definition:

“In addition to the high intensity combat operations that are traditionally associated with military operations, the 21st century mission space has expanded to include a wide spectrum of mission challenges, ranging from providing support to multi-agency disaster
Complex endeavors are operations and challenges which, due to their very nature, do not adhere well to hierarchical and deconflicted approaches to C2. An argument has been made that to tackle the challenge of complex endeavors, more networked C2 approaches are required. However, while it may be true that networked C2 approaches are more well-adapted to manage complex endeavors, it has also become evident that there is no “one size fits all” approach to C2. One aspect of the uncertainty associated with complex endeavors is that they are dynamic. Therefore, the circumstances constituting the endeavor may change over time. Successful operation within a complex endeavor is therefore dependent on the C2 organization’s ability to remain agile, ensuring that performance is kept within acceptable parameters when faced with changes in operational circumstances. Essential to the ability of an entity to cope with change is its ability to adapt its C2 approach to fit the requirements imposed by those changes. This ability is referred to as C2 Agility (NATO SAS-085, 2013).

C2 Agility is related to two concepts, the Approach Space and the Endeavor Space. The Approach Space represents a map of all C2 approaches ranging from deconflicted hierarchies to decentralized and entirely networked edge organizations. The Endeavor Space represents the problem space of an operation and encapsulates all possible circumstances and characteristics an endeavor may exhibit. Therefore, C2 agility represents an entity’s ability to change its position within the C2 Approach Space in response to changes in the Endeavor Space.

The North Atlantic Treaty Organization (NATO) research task group SAS-143 is currently conducting research into multi-domain C2 operations. As part of their work, SAS-143 seeks to further the understanding of C2 Agility theory as it relates to the C2 domain. An essential aspect of this is to better understand how the Endeavor Space impacts multi-domain C2 operations. In previous research conducted by SAS-085 (2013)—and studies since—the Endeavor Space has been characterized in multiple ways without any systematic agreement in regards to its constituent parts. Therefore, to further the understanding of the Endeavor Space and its relation to C2, a subgroup
of SAS-143 has been tasked with creating a uniform and systematic characterization of the Endeavor Space (NATO SAS-143, 2018).

1.1 Purpose
The Swedish Defense Research Agency (FOI), in collaboration with NATO research group SAS-143, are conducting research into how entities perceive and assess complexity as a function of their organizational structure and the nature of the environmental context in which they operate. The contribution of the current study would be to spearhead this effort by conducting initial research on how the Endeavor Space dimensions may be implemented in a C2 simulation testbed called the Experimental Laboratory for Investigating Collaboration, Information-sharing, and Trust (ELICIT). The study would also serve the purpose of creating the initial version of a tool intended for measuring perceived Endeavor Space position. This effort would involve designing and creating ELICIT scenarios that would correspond to distinctly different regions in the Endeavor Space. These scenarios would then be used to perform human-in-the-loop experiment trials to determine whether the scenarios evoked measurably different subjective perceptions (as measured by the developed self-assessment tool) of task complexity. This study would further add to the body of ELICIT research and Endeavor Space theory.

1.2 Research questions
To guide our research, the following research questions were formulated:

RQ1. How can the Endeavor Space dimensions be instantiated and manipulated in the ELICIT simulator?

RQ2. How do manipulations of Endeavor Space dimensions manifest as subjective experiences?

RQ3. How do manipulations of Endeavor Space dimensions affect how individuals perceive difficulty, and how does it impact their performance?

The relationship between these research questions is somewhat circular. Only by successfully measuring the perceived effects of dimension manipulations can it be
determined whether the ELICIT instantiations represent reliable manipulations of the Endeavor Space dimensions.

1.3 Hypotheses

Based on the available literature, several hypotheses were developed:

H1) Perceived low tractability is associated with an increase in subjective difficulty.

H2) Perceived high dynamics is associated with an increase in subjective difficulty.

H3) Perceived high dependencies is associated with an increase in subjective difficulty.

H4) Perceived low tractability is associated with a lower performance score.

H5) Perceived high dynamics is associated with a lower performance score.

H6) Perceived high dependencies is associated with a performance score.

Answering these hypotheses would serve to gain insight into whether the Endeavor Space dimensions were successfully instantiated and manipulated in ELICIT.

1.4 Delimitations

Although the C2 Approach Space include several organizational configurations (coordinated, collaborative, edge), this study was limited to include only a de-conflicted approach (in a hierarchical setup) due to the time and personnel requirements that would be needed to test each and every C2 approach. This vast body of work is rather suited for agent-based simulations. Furthermore, the purpose of the current research was to study dimensions of complexity, not how the choice of C2 approach affects how complexity is perceived.
2 Theory

In this chapter we will introduce the reader to the theoretical and literary background necessary to properly understand the purpose and contributions of the current study. We will begin by explaining the command and control (C2) concept.

2.1 Command and Control (C2)

“Command and control (C2) refer to the set of organizational and technical attributes and processes by which an enterprise marshals and employs human, physical, and informational resources to solve problems to accomplish missions.” (Vassiliou, Alberts, & Agre, 2015, p. 1).

The definition cited above is rather general in its purpose, but it encapsulates the core of what constitutes command and control. In its classical sense C2 often refers to a hierarchical organization tasked with the completion of a mission, e.g. winning a battle or managing a crisis. This is accomplished through two functions; the Command function and the Control function. It is worth noting that different definitions of C2 offer different emphasis on these two functions. Some definitions are more focused on control while others emphasize command, although the two concepts refer to different functions and processes, they are interconnected (Alberts & Hayes, 2006; Pigeau & McCann, 2002; Shattuck & Woods, 2000). We will begin with describing command and control as separate entities, then move into how these are interconnected in a system view. Lastly, we provide some examples of different C2 approaches that exist today.

Pigeau and McCann (2002, p. 3) define Command as the following: “Command is the creative expression of human will necessary to accomplish a mission.” Although the above definition is taken from a military context, it does capture the role and purpose of the command function in most C2 situations. The purpose of Command is to establish the operational intent from which a unified goal can be established and propagated through the C2 enterprise. This unified goal provides focus and acts as an instruction for the control function, which is primarily concerned with the management of resources in pursuit of accomplishing the goal provided by command (Alberts & Hayes, 2006; Salmon et al., 2008). The purpose of command is not micromanagement but to broadly approach a problem and decide what must be done
to solve it; which objectives must be accomplished, what resources are needed, and to some degree decide how to deploy those resources. What can be considered mission success or failure in C2 is multifaceted issue, but the failure of command to properly communicate and enforce the commander’s intent throughout the organization is one factor associated with mission failure (Alberts & Hayes, 2006; Builder, Bankes, & Nordin, 1999).

Continuing with Pigeau and McCann’s definition, “control refers to those structures and processes devised by command to enable it and manage risk” (2002, p. 4). Control includes all logistical, informational, and tactical processes and structures required for facilitating mission success. These include the structure of the organization in place for managing the operation, structures for monitoring and assessing mission progress, rules and constraints e.g. ROE (rules of engagement and schedules), the organization’s distribution of decision rights and responsibilities, etc. (Alberts & Hayes, 2006; Pigeau & McCann, 2002). The control function and the control systems it deploys is primarily intended to reduce uncertainty and render the problem space manageable for command. While not all uncertainty can be eliminated, the processing, structuring, and management of information regarding own resources, the operational environment, and the enemy makes the problem space more amenable to intervention. The reduction of uncertainty provided by the control functions enables more time to be committed to producing faster and more effective courses of action (McCann & Pigeau, 1998). The effort of grounding a problem space through procedure and process does however come at the cost of an organization’s flexibility. In complex and dynamic environments—such as military environments—it is important to consider in what ways adopting control limits an organization’s ability to adapt to different situations (Brehmer, 2008; Pigeau & McCann, 2002).

Although the above description paints Command and Control as two distinct functions they do—as previously mentioned—interact and one is not possible without the other. Most literature around C2 therefore view C2 from a system perspective. For context it is therefore important that we provide a brief overview of system theory and its relation to C2.

The study of systems, e.g. in fields like general system theory (Von Bertalanffy, 1968), cybernetics (Ashby, 1957), or joint cognitive systems (Hollnagel & Woods, 2005), refers
to the study of entities which are composed of several interrelated and interdependent elements that through interaction with each other create a whole that is greater than the sum of its parts. The study of systems is thus concerned with the study of wholeness, often in domains where the study of the isolated part is not enough to understand the overall system behavior (Von Bertalanffy, 1968). In the case of C2 systems this references how the functional parts—the people, the technology, the processes and structures, etc.—constitute the functional whole, the goal-directed C2 system (Brehmer, 2008; Salmon et al., 2008).

One of the dominant perspectives for modelling C2 systems stem from the field of cybernetics. Weiner defined cybernetics as “the science of control and communication, in the animal and machine” as quoted in Ashby (1957, p. 1). In relation to C2 this entails the modelling of the process of command and control, its components, their order, and the C2 systems interaction with its environment. One such model was presented by Lawson (1981)—illustrated in Figure 1 below—capturing the C2 as a process of sensing the environment, processing and comparing the information gathered towards a desired state in order to develop a basis for making a decision on how to act. This process forms a loop as the environment changes due to the impact of the chosen course of action, and a new evaluation must be made.

![Figure 1. Lawson’s model, adapted from Lawson, 1981.](image)
The cybernetic models of C2 have however been criticized for being too focused on modelling the control process, neglecting essential aspects of the C2 system. Two main critiques are that 1) cybernetic models discount the initiative involved in command, thus characterizing C2 as only being reactive in the face of change unable to act with initiative (Builder et al., 1999), 2) cybernetic models lack several functions essential to a functioning C2 systems (Brehmer, 2005). In order to address these faults, Brehmer (2005, 2006) developed the D-OODA loop (see Figure 2 below), which, he argues, better captures the essence of C2 systems and the environment in which they operate.

![Figure 2. The D-OODA loop, adapted from Brehmer (2006).](image)

Similarly to Lawson’s model, the D-OODA loop models C2 as a system which continuously senses the environment (sensing and processing) in order to generate a course of action (compare and decide) that is aligned both with the overall mission goal (the desired state) as well as the current operational conditions (the environment) (Brehmer, 2006). The critiques towards Lawson’s model raised earlier are in the D-OODA loop remedied through a more correct representation of a C2 system’s functions—encapsulated on the left-hand side of the model in figure 2—as well as an explicit description of these functions and how they enable the C2 process. This allows the D-OODA model to more closely represent a C2 system as a dynamic process.
capable of adaptation compared to the static process represented in Lawson’s model (Brehmer, 2005, 2006). The three main functions enabling C2 according to the D-OODA model are the information collection function, sensemaking function and planning function. The information collection function monitors the environment through various sensors and converts the collected data into actionable information for the sensemaking function. The planning function produces orders based on the directives provided by the sensemaking function. Central to both planning and information collection is the sensemaking function. Sensemaking refers to the function tasked with establishing the mission directive most suitable for the current operational situation. The sensemaking function decides what needs to be done, this decision is made in light of the overall mission objective as well as an understanding of the current situation. This is illustrated in Figure 2 as the sensemaking function taking as input the mission—the overall mission statement—and information from the information collection function. The bidirectional relationship between information collection and sensemaking indicate that while the information collection function provides sensemaking with information, sensemaking also provide instruction regarding which information must be collected by the information collection function to satisfy the sensemaking function’s information needs (Brehmer, 2006).

One final perspective necessary to account for C2 as a system is C2 as a sociotechnical system. Sociotechnical system theory refers to the study of systems in which social systems (e.g. people, groups and organizational structures) are interrelated with technical systems (e.g. communication systems, sensor networks etc.) (Walker, Stanton, Salmon, & Jenkins, 2009). These are systems in which the conditions for success or failure are born of the interactions enabled by the linear nature of technology and the nonlinear nature of human factors. The focus in the study of sociotechnical system thus lies with achieving increased system performance through joint optimization, meaning that to successfully create and operate sociotechnical systems there must be a balance between social and technical factors. As these two factors are interrelated, a failure to balance the two would lead to an increased quantity of uncontrolled and undesired interaction which may detrimentally impact system performance. The goal of joint optimization is to create organizations which can cope with circumstances that display open system properties such as complexity and
C2 is regarded as a sociotechnical system due to its functions being dependent on the successful interaction between human elements (the commanders, and operators organized within the system) and the technical aspects (e.g. sensor networks, weapon systems etc.) which constitute today's C2 systems. Thus, a C2 system’s organizational structure—together with the technologies it has at its disposal—will yield different capabilities depending on how these two factors are allowed to interact. For example, this stands at the core of concepts such as network enabled capability (NEC), where a decentralized approach to command (organizational factor) coupled with the powerful information distribution capabilities afforded by new technologies yield the possibility to create a network-enabled organization that do not need to rely on hierarchical chains of command, and is therefore argued to be better at managing complex and dynamic environments (Walker et al., 2009).

To understand the direction in which modelling of C2 and C2 theory has developed it is important that we shortly account for the classical view on C2 and the issues related to it. Classical C2—also referred to as industrial age C2 due to its development during the industrial age—is characterized by its hierarchical and layered organizations in which the larger problem (the mission) is deconstructed into smaller subproblems to allow specialized sub-entities to solve each subproblem separately. The approach is thus largely dependent on the ability to deconstruct a problem to fit it into the organization's problem-solving hierarchy and optimizing each entity’s specialized subcomponents ability to solve their assigned subproblems (Alberts & Hayes, 2003, 2007).

Similar to the cybernetics models, the issue with this perspective is that it suggests a “one size fits all” approach to C2. From this perspective, the effort of optimizing C2 is focused on optimizing the problem-solving capabilities of the individual sub-entity (Alberts & Hayes, 2003; Builder et al., 1999). The reality of information age operations is however that the “one size fits all” approach is no longer applicable (Alberts & Hayes, 2003, 2007; NATO SAS-085, 2013). Moving into the information age, C2 systems are facing a far more diverse set of complex challenges than those faced in the industrial age. These endeavors are characterized by often requiring involvement from multiple different organizations (not just military) to be solved, not being amenable to
breakdown into subproblems, hard to predict, and thus difficult to control (Brehmer, 2006; NATO SAS-085, 2013).

To rise to the challenges posed by this new age, C2 entities must evolve the way in which problems are approached. Different organizational approaches will be better suited for coping with different problems, and therefore entities that have the ability to adapt their organization to the problem at hand will be the most successful in handling these complex endeavors.

2.1.1 C2 Approach Space

As was explained in section 2.1, there are different ways in which C2 may be approached. An example was given of what is considered the classical approach to C2 characterized by hierarchical organizations with centralized command, deconflicted units, and tightly controlled processes (Alberts & Hayes, 2003). Such approaches are strongly contrasted by network-centric approaches which advocate robustly networked organizations with decentralized command and autonomous units (Alberts & Hayes, 2003). Besides hierarchical industrial age C2 and networked information age C2 there are multiple approach variants which have been adopted by several different organizations throughout history. In the debate on which approach is the optimal one, one central issue has been that although there has been much debate on the topic there has existed no aggregated model in which all C2 approaches can be captured and studied for comparison (Alberts, Chan, Bernier, & Manso, 2013).

In an effort to remedy this, the NATO SAS-50 research group endeavored to develop a means of modelling C2 approaches that would make them amenable to systematic categorization. The result of this effort was the C2 Approach Space (NATO SAS-050, 2007), illustrated in Figure 3. The C2 Approach Space is a conceptual model intended to act as a representation of the option space of available C2 approaches. However, the Approach Space is not only intended to designate each C2 approach a position within the approach space; Each position must represent the different set of behaviors associated with the approach inhabiting that position so that conclusions may be drawn regarding differences in effectiveness between the approaches (Alberts et al., 2013). The Approach Space is modelled along three dimensions which are central to
the characterization of a C2 approach: the *allocation of decision rights*, the *patterns of interaction*, and the *distribution of information*.

![Diagram of the Approach Space](image)

**Figure 3. The Approach Space. Source: SAS-085 (2013).**

Allocation of decision rights (ADR) refers to how an organization chooses to assign authority and responsibility across the organization. It may vary between being allocated entirely to one actor or distributed equally across the organization. This may be decided both explicitly and implicitly through the practices and rules of an organization, however the ADR may also be influenced through emergent behaviors (Alberts et al., 2010; NATO SAS-050, 2007).

Patterns of interaction (PoI) concern the ability and willingness of actors to interact and create interaction. This ability is influenced by the organization’s information structure, the degree to which actors are encouraged to cooperate, and their ability to create interaction and cooperate with other actors. PoI may be tightly constrained, as seen in typical hierarchical organizations, or unconstrained (Alberts et al., 2010; NATO SAS-050, 2007).
Lastly, the distribution of information (DoI) addresses the way in which information is distributed across the organization. Much like PoI, this may vary between being broadly distributed across the organization and having no distribution of information (NATO SAS-050, 2007).

Although the dimensions are presented separately, they are not to be viewed as entirely independent from each other; they are interdependent in nature and changes made to one dimension may have consequences for the other two. For example, broadening the ADR immediately influences the PoI possible within the organization and the DoI required to sustain this. Therefore, changes made along either dimension should be followed with appropriate changes to the other dimensions to avoid the organization becoming dysfunctional (NATO SAS-050, 2007; NATO SAS-085, 2013). As C2 is often practiced in unstable environments where plans and processes may shift as the situation develops, an organization’s actual position along these dimensions may be different from the intended position. Therefore, a central aspect of the C2 Approach Space is that it must capture a system’s actual position rather than the position that is intended in theory (Alberts et al., 2013; Bernier, Chan, Alberts, & Pearce, 2013; NATO SAS-050, 2007). Besides situational factors, an organization’s ability to adopt a certain approach may also be influenced by factors such as the organization’s doctrine, culture, capabilities, and available resources (NATO SAS-050, 2007; NATO SAS-085, 2013).
Figure 4. The Approach Space regions. Source: Alberts et al. (2010).

Although the specific position of an adopted approach may vary from its intended location, the characteristics of that C2 approach will be generally associated with a specific region of the Approach Space (Alberts et al., 2013). These regions can be classified as the following: Conflicted C2 approaches, De-conflicted C2 approaches, Coordinated C2 approaches, Collaborative C2 approaches, and Edge C2 approaches (Alberts et al., 2010). These approach regions are placed along a diagonal line stretching from the corner of the conflicted C2 region to the corner of the Edge C2 region of the Approach Space (see Figure 4). The regions indicate that C2 approaches may still differ in their precise position but given that the approach display certain characteristics it will generally occupy a position within a corresponding region (Alberts et al., 2013). Although the C2 Approach Space itself makes no such assumptions, a commonly held belief in C2 research has been that C2 approaches located closer to the Edge region of the Approach Space are more adaptable and thus better at tackling complex endeavors. However, more recent C2 research has concluded that it is an entity’s ability to appropriately match and adapt their C2
approach to the context of operational challenges that is the most important ability (Alberts et al., 2010).

Next, we shall review the efforts made in C2 research regarding this ability to adapt the organizations approach to the context of operational challenges.

2.2 C2 Agility

The endeavors facing the C2 entities of today has drastically changed from the more monotonous challenges of the industrial age. Today’s C2 organizations must face the reality of working in complex mission environments that require complex enterprises to tackle the challenges. Whilst previously a military organization would only handle military matters, it may today be involved in problems related to politics, social matters, and infrastructure (Brehmer, 2006; NATO SAS-085, 2013). This has created a much deeper and more complex range of circumstances that military organizations may be faced with. This change in scenery has required a departure from the “one size fits all” mentality of old (in terms of organizational structure), a departure which has triggered a re-evaluation of the capabilities an entity requires to be successful in such complex environments. One capability identified to be beneficial to the success of an entity operating in a complex mission environment is agility (NATO SAS-085, 2013).

In terms of C2 this entails the ability of a C2 entity to be responsive and flexible in regard to what type of C2 approach to apply when facing unpredictable and unstable circumstances.

NATO research group SAS-085 defines C2 agility as “The capability of C2 to successfully effect, cope with, and/or exploit changes in circumstances....C2 Agility enables entities to effectively and efficiently employ the resources they have in a timely manner” (NATO SAS-085, 2013, p. 20). C2 agility is thus viewed both as a outcome and a capability. Agility can be an outcome in the sense that an entity may either succeed or fail at manifesting agility. Agility as a capability refers to an entity’s inherent potential to manifest agility based on an understanding of which attributes and behaviors enable or inhibit agility (NATO SAS-085, 2013).

The core prerequisite to agility is the presence of change, either as change that is detrimental and poses a threat to the endeavor or change that presents an opportunity
to succeed. Without change, the principles of agility do not apply (NATO SAS-085, 2013). Thus, measuring and observing agility—or the lack thereof—is a question of examining the degree to which an entity is able to respond to change. The ability to respond to a change will be a function of the entity’s ability to detect the change, how the entity is affected by the change, and how well it responds to the change. To determine how well an entity copes with the change, a baseline is required that represent what would have happened if the change had not occurred. This may then be compared to the entity’s actual performance in regard to detecting the change, deciding on a course of action, implementing the action, and whether or not the action has the desired effect. If the entity is able to keep performance within acceptable bounds it has manifested agility. If not, it is lacking in agile capability. If the time between the detection of a change and the implementation of an appropriate mitigating response action is too long, it may be an example of the entity lacking agility due to being insufficiently responsive. An illustration of such an example is shown in Figure 5 below (NATO SAS-085, 2013).

Figure 5. Example of how lack of agility can be observed. Source: NATO SAS-085 (2013).

Agility in terms of C2 is a question of how well a C2 system is able to accomplish its functions across a range of missions and circumstances, i.e. across the endeavor space. How well a C2 system is able to execute its functions depends on the C2 approach that is being used in the context of the endeavor that the C2 system is engaged in. Therefore,
in a complex mission environment, the success of a C2 system lies in its ability to maintain effective execution of its functions despite changes in circumstances. This would require the C2 system to be able to adapt its C2 approach to the requirements imposed by the current circumstances of the endeavor, in order to ensure that the C2 system is functioning effectively across the entire endeavor space. A C2 system that is able to do this is considered fully agile. Utilizing the wrong approach may cause immediate mission failure or detrimental performance in vital phases of an operation, resulting in mission failure (NATO SAS-085, 2013). For example, in the graph depicted in Figure 5 above, it can be argued that the lack of responsiveness in the organization could be due to a mismatch between the endeavor and the C2 approach employed by the organization. Agility in a C2 organization is determined by two factors: its C2 approach agility, which is the repertoire of C2 approaches the organization has at its disposal; and its C2 maneuver agility - the organization's ability to move between the C2 approaches available to it. C2 approach agility requires an entity to obtain and maintain a number of different C2 approaches that may be employed depending on the state of the endeavor space. Agility in regards to maneuverability refers to an entity's ability to identify change and recognize the effects said change has on the viability of the currently employed C2 approach, it also requires the entity to identify which C2 approach in its toolkit is more appropriate given the circumstances and adapt its approach within an appropriate time frame (NATO SAS-085, 2013).

2.2.1 C2 Endeavour Space

The aforementioned ability (see section 2.2) of an entity to adapt their C2 approach given the current operational challenge would require a understanding of which C2 approach is best suited for dealing with a certain set of mission circumstances. This could be achieved by mapping the Approach Space to a model that similarly divided the problem space into regions. This is the purpose of the Endeavour Space (NATO SAS-085, 2013).

Similar to how the Approach Space seeks to aggregate all C2 approaches and map these to regions in a common space, the Endeavor Space seeks to encapsulate all possible missions and circumstances a C2 system may face (NATO SAS-085, 2013). The necessity of this lies in the need for being able to identify under which operational
circumstances a C2 system is operating so that conclusions can be drawn regarding the appropriateness of the C2 approach being used (Johansson, Carlerby, & Alberts, 2018). Figure 6, below, illustrate this relationship. Previous research—e.g. NATO SAS-050 (2007), NATO SAS-085 (2013), and Alberts et al. (2010)—has highlighted the need for an Endeavor Space and established the need for mapping the Approach Space regions to corresponding regions in the Endeavor Space. However, no extensive research has yet been focused at explicitly modelling the Endeavor Space.

Figure 6. Mapping between a C2 Approach Space region and a Endeavor Space region, illustrated from Johansson et al. (2018).

One of the first attempts at conceptualizing the Endeavor Space was made by Johansson, Carlerby, and Alberts (2018), who postulate that the Endeavor Space should be viewed as “a system with certain properties that affect the appropriateness of a given C2 approach” (Johansson et al., 2018, p. 4). Johansson and colleagues suggest that the Endeavor Space should be framed according to three dimensions: Coupling/Causality, Dynamics, and Degree of Complexity/Tractability (Johansson et al., 2018).

Coupling/Causality refers to the level of interdependence of the components or entities that constitute a particular problem. This may vary between strong interdependence or weak interdependence (Johansson et al., 2018). The coupling of a system determines the kind of interactions that may take place and thus the types of consequences interactions may have. In a loosely coupled system where components are not interacting, effects and disruptions may be more isolated, but it may be more difficult to investigate why something has happened due to lack of causal relationships between components. Conversely, in a tightly coupled system it is easier for interactions between components to occur and effects may propagate through the
system faster and have widespread consequences. The degree of coupling between components also influences the level of determinism between cause and effect.

Dynamics is concerned with the potential rate of change as well as the amplitude of change inherent in the system. A system may range between being highly dynamic or being low in dynamic behavior (Johansson et al., 2018). Dynamics is a product of interaction and encapsulates both the potential for an entity to affect its environment and the potential ways in which the environment can affect the entity (Feibelman & Friend, 1969; Jensen & Brehmer, 2003). Thus, dynamics is also a factor related to time pressure, as a dynamic system may change in real-time and therefore increases pressure to make decisions at the right time (Jensen & Brehmer, 2003). A system's potential for change also determines the potential of being surprised, as dynamics increases so does the potential that surprising events will occur that may force redirection of strategy (Johansson et al., 2018).

Finally, degree of complexity/tractability refers to the degree to which it is possible to describe and understand what is happening within the system. This may also describe the potential for surprising and undesired events that may occur in the system. A problem may range between being intractable and tractable (Johansson et al., 2018). The complexity of a problem is directly associated with an observer's capacity to understand it. If a problem is said to be complex it should be difficult—or even impossible—to identify cause-and-effect relationships and predict outcomes (NATO SAS-085, 2013). Complexity is thus also viewed as a component of difficulty (Braarud & Kirwan, 2011). Therefore, complexity carries a subjective component in that the complexity of a system is—in part—a product of the domain-knowledge and experience of the observer or agent interacting with the system (Braarud & Kirwan, 2011; Haerem, Pentland, & Miller, 2015; McIntyre, 1998).

The dimensions refer to characteristic features of the problem space that may have an impact on the appropriateness of a chosen C2 approach. For instance, Edge approaches may be better suited for managing endeavors that are highly dynamic, with components that are loosely coupled, and where the problem is difficult to grasp (i.e. intractable). Conversely, a less networked approach may be more appropriate when dealing with an endeavor that is more tractable due to being more static and having strongly coupled elements (Johansson et al., 2018).
It should be noted that the Endeavor Space dimensions are rather generically formulated. This is however a necessity owed to the purpose of the Endeavor Space to encapsulate a vast number of potential operational circumstances. A too strict and narrow definition of these dimensions would exclude certain environments and thus hamper the Endeavor Space’s generalizability.

2.3 Simulations

The English noun *simulation* stems from the Latin verb *simulare*, meaning “to copy, represent, feign” (“Simulate,” 2019a; “Simulation,” 2019). Synonyms and related words of the English verb *simulate* include, for example; *pretend, impersonate* and *imitate* (“Simulate,” 2019b). The concept of simulation, in scientific contexts, refers to the act of constructing a representation that mimics, imitates, or replicates operations and core behaviors of a target system (Grüne-Yanoff & Weirich, 2010; Laurids Boring, 2011). A system can be defined as a collection of component parts—e.g. people, objects, hardware, software, policies, etc.—interacting in such a way that their collective overall performance outcome is unachievable by individual system components (Banks, 2010). These components interact in different ways to generate system-level behavior, performance, and phenomena (Banks, 2010; Grüne-Yanoff & Weirich, 2010). There are several reasons why simulations are needed. For instance, the target system or phenomena of interest may be inaccessible (e.g. distant star systems or black holes), too dangerous to set up and practice (e.g. space missions), prohibitively expensive to build and test (e.g. nuclear power plants), morally or ethically unacceptable (e.g. medical trials or the dynamics of epidemic outbreaks), any combination of these factors, or may not even exist at all (Banks, 2009; Hollnagel, 2011; Moroney & Lilienthal, 2009). Simulations, then, are tools and techniques that allow for accessible, safe, and repeatable observations, interactions, and analyses of complex systems and the emergent behaviors or phenomena they generate.

In the following subsections, we will describe some central concepts and definitions of simulations, explain how simulations can be categorized, and discuss how simulations are applied for different purposes.
2.3.1 Concepts and Definitions

In the literature, two concepts sit at the very center; *models and simulations*. Broadly speaking, models are typically described as approximate representations of reality, and simulations are implementations—or executions—of models that allows for repeated observations and analysis of the underlying model (Banks, 2009, 2010; Grüne-Yanoff & Weirich, 2010). However, the word *model* has also been used to refer to physical objects that in turn simulate different phenomena. For example, Grüne-Yanoff and Weirich (2010, p. 22) describe a scale model of San Francisco Bay that uses hydraulic pumps to “simulate the action of tidal and river flows in the bay, modelling tides, currents, and the salinity barrier where fresh and saltwater meet.” Here, *model* is used both as a noun (a scale model) and a verb (modelling tides and currents). The subject (the scale model) is described as being an actor of sorts, simultaneously performing the acts of simulating tidal and river flows, and modelling tides and currents. Similarly, Schlimm (2009) describes how neurophysiologist William Grey Walter invented autonomous robots whose behavior were generated by only two functional parts - analogue neurons. Each robot consisted of two “neurons” along with motors, batteries, relays, wheels, and photoelectric cells. Grey Walter’s idea was that it is the interconnectivity of neurons, rather than the number of neurons themselves, that generate complex behavior. Despite their simple construction, and the fact that they only consisted of two mechanical neurons, the robot “tortoises” displayed remarkably complex and animal-like behavior, such as navigating around obstacles while moving towards a source of light (Schlimm, 2009). Similar to how Grüne-Yanoff and Weirich (2010) refers to the San Francisco Bay as a model, Schlimm (2009) describes Grey Walter’s “tortoises” as models, yet they are also described as simulations of animal behavior. Schrage (1999) describes models as high-level abstract representations of reality that can take the form of equations scribbled on a paper, full-scale models of passenger jet aircraft, and anything in between. Additionally, simulations are described as “virtual models of processes” and physical models are referred to as “prototypes” (Schrage, 1999, p. 7). Although not entirely interchangeable concepts, models, simulations, and prototypes have become—from a practical point of view—synonymous in the sense that they are all meant to be meaningful re-creations of some selected portions of reality (Schrage, 1999).
Simuland, Referent, Model, Simulator

In his doctoral thesis, Rybing (2018) presents a conceptual overview to explain the terminology associated with simulation. It contains four main concepts: the simuland, the referent, the model, and the simulator (see Figure 7).

![Diagram of the relationship between simuland, referent, model, and simulator](image)

*Figure 7. Central concepts of simulation and how they relate to each other. Adapted from Rybing, 2018.*

The simuland represents the real-world objects that are to be simulated and studied, e.g. physical objects, processes, or phenomena, along with any relevant factors that may act as external forces on the primary target system (Rybing, 2018). The referent is any available quantitative knowledge—e.g. equations or statistical data—and qualitative knowledge—e.g. theories, observation data, or personal experience of the simuland—about the simuland that the researcher can draw on when selecting what features and characteristics of the simuland to model, and design and construct the simulator with which to execute (simulate) the model (Liu, Macchiarella, & Vincenzi, 2009; Rybing, 2018).

The model is a selective, simplified, and abstract representation of the simuland—or rather the referent (Rybing, 2018)—and can be e.g. physical, mathematical, or otherwise logical in nature (Banks, 2010; Moroney & Lilienthal, 2009; Rybing, 2018). However, whereas scientific theories should ideally be accurate and true regarding whatever they are intended to explain, models need not be (Rybing, 2018). Rather,
models can be viewed as mediators between reality and theory (Grüne-Yanoff & Weirich, 2010). They are tools whose purpose is to help us understand and explain the world through new and more accurate theories. As previously mentioned, models are—amongst other things—simplified representations of real-world objects or entities. By necessity, models must be selective and simplified representations as opposed to exhaustive, detailed, and fully authentic re-creations of the simuland. Otherwise they lose their explanatory power. Brehmer referred to this as the “cat problem” where the best model of a cat is another cat. But if the second cat were a perfectly authentic model of the first cat it would be just as perplexing and inexplicable as the original cat, which defeats the purpose of modelling the original cat in the first place (Brehmer, 2008).

The intended purpose of the model typically influences the modelling process where features of the simuland—via the referent—are selected for model inclusion (Petty, 2009).

The fourth component in Rybing’s model is the simulator, which is the setting, device, computer program, or system that implements and executes the model over time to generate or produce the simulation (Rybing, 2018). However, the line between the simulator and the simulation is easily blurred. As Rybing explains, computer-based simulations consist of the hardware (the simulator) executing the software (the model), but in cases where human participants act out roles in e.g. war-games or trauma care training exercises the simulator is not as easily separated from the simulation. In such cases, the argument can be made that the people partaking in the activity are the simulator and that they are implementing and executing a shared mental model of the “real” system or situation they are enacting (Rybing, 2018).

To summarize these four basic concepts, the simuland is the real-world systems and objects to be studied, the referent is the available knowledge about those systems and objects. A model of the simuland is created by simplifying and abstracting select portions of the referent. The model is then executed by some device or system (the simulator) to generate a simulation behavior that can be repeatedly observed and studied to learn about the simuland (Rybing, 2018). It is worth noting, however, that this definition of a simulation—as a product of a model somehow being executed—is not universally agreed upon. While it is a common definition to describe computer-based simulation (Banks, 2010; Frigg & Reiss, 2009; Grüne-Yanoff & Weirich, 2010;
Rybing, 2018) where the model is often mathematical, others argue that the intended purpose of a simulation is part of its definition (Rybing, 2018). Still others argue that a simulation could be considered as such as long as it its behavior is believed to be similar enough to some other system that one could learn about the other system by studying the generated behavior output of the simulation (Rybing, 2018). Consider these three interpretations and compare how they relate to a weather forecast system, simulation-based training, and Grey Walter’s tortoise robots. The weather forecast system does indeed execute a (mathematical) model over time to predict and illustrate future weather states. In contrast, a simulator intended for use in training, such as a CPR mannequin, could be argued to execute a model as it is used by a human participant learner—as previously discussed—but the point of the simulator is not to explore or make predictions about the human cardiovascular or respiratory systems, but to teach the students how to perform proper chest compressions and artificial ventilation. Grey Walter’s tortoise robots could be considered to be simulations of animal behavior if we can learn something about animal behavior by studying the tortoise robots, even though they were not necessarily designed or intended to behave in any particular way since the fundamental simuland was actually a neuron rather than an animal. In this sense, what constitute a simulation is a rather philosophical and subjective notion (Rybing, 2018).

Validity

Another important concept in simulation is validity. As previously mentioned, models are not necessarily true. It thus follows that simulations are not necessarily true either, and should not be viewed through the same critical lens as theories (Rybing, 2018). However, much like how Box’s (1976) aphorism holds that “all models are wrong, but some are useful”, the same could be argued for simulations. When assessing the “usefulness” of a simulation one must consider two things: is it based on a good model, and is the model properly implemented in the simulator? The process of answering these questions is called validation, where one determines the quality or degree to which the simulation mimics the real-world system it was designed to simulate (Banks, 2010; Feinstein & Cannon, 2002; Moroney & Lilienthal, 2009). The validity of a simulation can be discussed in several ways. Generally speaking, validity concerns the accuracy with which a simulation represents the real world, i.e. the simuland (Banks,
Furthermore, this assessment—typically made by subject matter experts—is made against the backdrop of the intended use and purpose of the simulation (Banks, 2009, 2010; Moroney & Lilienthal, 2009; Rybing, 2018). For this reason, there are numerous different validity concepts that are of greater or lesser interest to the researcher. For the sake of brevity, we will not provide a list of these concepts here, however the interested reader is referred to Feinstein and Cannon (2002) or Rybing (2018) for further information about these.

Instead we will briefly explain how Feinstein and Cannon (2002) provide a framework to group these concepts together and explain how they relate to each other. They present four concepts; representational validity, educational validity, internal validity, and external validity, where the first two are relevant to the development and application of a simulation, and the last two relate to its structure and function (Feinstein & Cannon, 2002; Rybing, 2018). Representational validity is described as a process rather than an end goal, beginning at the earliest phases of design, modelling, and development of a simulation (Feinstein & Cannon, 2002). It refers to how representations in the simulation capture the conceptual “essence” of the simuland. It therefore deals with theoretical rather than literal accuracy. As such, a simulation need only to be adequately homomorphic (similar in structure) in relation to its simuland in order to be conceptually—and representationally—valid (Feinstein & Cannon, 2002). Analogous to how painters can capture and emphasize certain aspects and features of the real world (at the expense of others) to create paintings that somehow seem more real than actual photographs, modelers must “artistically” interpret and conceptualize key elements of the simuland, i.e. to capture its “essence” (Feinstein & Cannon, 2002).

If a simulation is developed for educational purposes, educational validity is of utmost importance. It is a measure not of the simulation itself but the effects it has on those who interact with it; does it stimulate learning by promoting positive transfer of training? It is an assessment of whether the simulation is fit for its educational purpose, and therefore relates to the application stage of simulation use (Feinstein & Cannon, 2002).

Internal validity addresses whether the simulation functions as it was intended and designed to. There are two components to internal validity. The first harkens back to the question of whether the correct things are simulated correctly. That is to say, have
the appropriate simplifications and abstractions been made when developing the model, and is the internal logic of the simulation structured in such a way that the instantiated model generates output that is coherent (Feinstein & Cannon, 2002; Rybing, 2018)? The second component of internal validity addresses the way participants and users can interact with, understand, and gain insights from the simulation (Feinstein & Cannon, 2002). This is different from educational validity in the sense that the participants need not gain new knowledge for the simulation to be internally valid (Feinstein & Cannon, 2002). However, internal validity could be considered a necessary—but not a sufficient—condition for educational validity.

Finally, external validity addresses the link between the simulation and the simuland, specifically whether the correct features were selected for simplification and abstraction in the modelling phase to begin with. It therefore determines—given that the simulation is internally valid—whether the generated output behavior reflects that of the simuland (Feinstein & Cannon, 2002; Rybing, 2018). For the purposes of simulation-based training, the relationship between educational and external validity is bidirectional in the sense that skills and knowledge taught in the simulation should be applicable in the real world, while students should also be able to apply their existing knowledge to enhance their performance in the simulation (Feinstein & Cannon, 2002; Rybing, 2018). Contrary to educational validity, however, external validity is concerned with replicating what would actually happen in the real world rather than teach any skills. An externally valid simulation would merely require and activate the same skills that are applied in the real world but would not necessarily have to provide students with any actual training (Feinstein & Cannon, 2002; Rybing, 2018).

To synthesize the previously described basic concepts of simulation with validity, Figure 8 shows how the intended purpose of a simulation affects the different components and illustrate how validity is related to each one.
Figure 8. The central concepts of simulation use and validation. The intended purpose of the simulation influences all stages of development and application. Adapted and modified from Rybing, 2018.

Fidelity

*Simulation fidelity* is often used as an umbrella term to describe the perceived degree to which the simulation resembles the real world (Banks, 2010; Liu et al., 2009). The quality of this similarity can be assessed on several fidelity dimensions, such as equipment fidelity, functional fidelity, environmental fidelity, task fidelity, or psychological and cognitive fidelity (Liu et al., 2009). However, no single accepted definition of the term exists (Liu et al., 2009). In the colloquial sense it refers to the level of physical realism of the simulation, i.e. whether it looks, sounds, feels, or even smells like the real thing (Liu et al., 2009). Thus, in a sense, fidelity is a property of the simulator and its behavior (Rybing, 2018).

Fidelity is most commonly associated with simulation-based training. The aviation industry has been heavily reliant on simulations for basic stick-and-rudder and instrument training, multicrew exercises, and other aspects of aviation training (Dahlstrom, Dekker, Van Winsen, & Nyce, 2009). Historically, it was assumed that higher simulation fidelity was always positive, but it has since been shown that the way a simulation interacts with and acknowledges the actions of the student is far more
important from a training perspective than the way the simulation looks (Dahlstrom et al., 2009; Liu et al., 2009). In fact, high (physical) fidelity can even have negative effects on transfer of training for novice students as they do not yet possess the skills or the experience necessary to look beyond the physical appearance of the simulation, and fail to focus on the fundamental mechanics they are supposed to practice and learn (Feinstein & Cannon, 2002). In this view—sometimes referred to as the “Alessi Hypothesis”—the relationship between fidelity and training effectiveness is nonlinear, and uniquely so for each student as other factors—such as experience or personal characteristics—affect the learning curve (Liu et al., 2009).

Unsurprisingly, the appropriate level of fidelity is also determined by the intended purpose of the simulation. Simulations for developing and testing theories may require higher levels of fidelity than do training simulations (Banks, 2010).

2.3.2 Simulation Typology

Simulations are typically classified into three categories: live, virtual, and constructive simulations (Andrews, Brown, Byrnes, Chang, & Hartman, 1998; Banks, 2010). Live simulations are when real people use and operate real equipment to perform real tasks. An example would be war games where military entities put real soldiers and weapon platforms in combat engagement situations—often with live ammunition—to let them experience “real” combat in order to prepare them for actual warfare in a way that only full-scale exercises can (Banks, 2010). Virtual simulations also involve real people but the systems they operate are simulated (Banks, 2010). Virtual simulations come in many forms as they utilize and combine physical objects and digital interfaces and environments. Aircraft cockpit simulators, for instance, involve an operating environment identical to the actual aircraft—its seat, sticks, buttons, and screens—and digital elements such as screens in place of the aircraft canopy to simulate the natural environment, and simulated targets and hit indicators on flight instruments and screens (Banks, 2010). Other examples of virtual simulations include virtual and augmented reality (VR and AR), part-task training simulators such as medical mannequins, or microworlds (e.g. C3Fire; Granlund, 2002). Virtual simulation is likely the most common class of simulations (Rybing, 2018). This is because they are primarily used for training which is a major application and use of simulations.
Live and virtual simulations are both common in human-centered simulation where human actions and behaviors are the core interests (e.g. role-playing or part-task training), but whereas human-centered simulations may be analogue or digital (or a combination of the two), human-in-the-loop simulations put an emphasis on computer-based simulation where human operators directly control functions of the digital system (Rybing, 2018).

The third type of simulations is constructive simulations where all elements of the system—i.e. people, equipment, environments, etc.—are simulated autonomously (Banks, 2010). Real people merely provide configurations for the initial input parameters, but they have no determinative impact on the outcome of the simulation itself (Banks, 2010; Rybing, 2018). The three most common examples of constructive simulations are: discrete-event simulations, commonly used to predict queueing behaviors in stores, traffic, or production lines (Diaz & Behr, 2010); continuous systems simulations, which uses differential equations to simulate physical systems (e.g. planetary systems or predator-prey population ratios) as they develop over time (Colley, 2010); and Monte Carlo simulations which aggregate repeated probabilistic calculations based on random selections from a range of possible input values. Monte Carlo simulations can be used to, for instance, approximate the value of π or estimate future earnings based on sales history (Sokolowski, 2010). Additionally, Andrews et al. (1998) argue for the existence of a fourth class of simulations in which elements of live, virtual, or constructive simulations are combined in hybrid simulations.

2.3.3 ELICIT

In order to study team performance dynamics in different C2 approach configurations, the United States Department of Defense Command and Control Research Program (CCRP, www.dodccrp.org) of the Office of the Assistant Secretary of Defense for Networks and Information Integration (OASD/NII) sponsored the development of ELICIT - an experimental research platform for studying how entities operate given (1) their C2 approach, (2) their team and individual characteristics, and (3) their network-centric capabilities, such as their ability to cooperate, their level of shared awareness, their synchronization, and their overall effectiveness (CCRP, 2010; Ruddy, 2007). In ELICIT, teams are tasked with uncovering the Who (color names, e.g. Blue group), What (e.g. embassy, financial institution, individual), Where (Greek letter nations, e.g.
Alphaland), and  *When* (month, date, hour, AM/PM) of an impending, fictitious, terrorist attack. To solve this problem, teams are organized—as hierarchies, edge organizations, or anything in-between—and individual team members are tasked with solving one or more of the four sub-problems. Pieces of information—called *factoids*—are distributed to the teams in waves at given time intervals. Factoids are one-sentence pieces of information, e.g. “all members of the Green group are in custody”, the consequence being that the Green group is *not* the *Who*, or “the Lion [the person planning and ordering the attack] is planning something on the 15th”, which would hint to the date-component of the *When* problem. Scenarios consist of 1-68 factoids in total. There are four categories of factoids: expert information, key information, useless noise, and support factoids that provide further contextual information. Team members must determine the relevance of the information and decide whether to pass it along to another member, and indeed *which* member, depending on which sub-problem the factoid is connected to. This makes the counterterrorism operation highly information intensive as individuals are given insufficient information to solve their problem(s) (CCRP, 2010; Ruddy, 2007). To solve a problem, several factoids must be collected; they are pieces in a logical puzzle. Thus, cooperation and information-sharing are crucial for the team to be successful. Therein lies the challenge.

The general user interface of ELICIT is depicted in Figure 9, below. It features an inbox where factoids appear as they are delivered by the game or sent by other team members. Factoids can be saved to the *My Factoids* list (see currently selected in Figure 9) by selecting them and pressing *Add to my Factoids*. 
Figure 9. The ELICIT interface. Factoids sent by the game or other members appear in the Inbox area (top). Saved factoids appear in the “My Factoids” list (bottom). The “Add to My Factoids”, “Share”, “Post”, and “Identify” buttons are located at the top of the screen. Buttons for accessing “My Factoids”, “How I’m Seen”, “What I See”, and the problem websites are located under the Inbox area of the screen.

Members can send selected factoids directly to each other—depending on their organizational structure—using the Share function (see Figure 10). They can also upload factoids to commonly available websites using the Post function (see Figure 11).

Figure 10. Peer-to-peer factoid sharing. Figure 11. Posting factoids to websites.

Each problem has its own corresponding website. Figure 12 depicts the website for the Who problem. Websites are only accessible to those who have been assigned the given problem.
Participants can get an overview of their own roles and the organizational structure of their teams using the *How I'm Seen* and the *What I See* functions (see Figure 13 and Figure 14), respectively.

![Figure 12. The Who website. Posted factoids are shown beside the name of the uploader.](image)

![Figure 13. How I'm Seen. Participants can see their assigned role within their organization.](image)

![Figure 14. What I See. Participants can get an overview of other organization members and their roles.](image)

Finally, when members believe they have solved their assigned problems they can use the *Identify* function (see Figure 15) to submit their answers. Responses are entered.
manually in plain text. The maximum allowed number of identify attempts is a configurable setting in ELICIT.

![Identify Screen](image.png)

**Figure 15.** The Identify screen. Responses to each question can be submitted at once or separately as their solutions are found.

Performance data—e.g. accuracy, number and direction of factoid transactions, number of attempted identifies, etc.—are recorded and stored in log files. These can then be analyzed to calculate overall team performance, shared awareness, quality of information sharing, and other useful metrics.

ELICIT is intended to be a platform for research and experimentation, analysis, and teaching on the information, cognitive, and social effects—e.g. information sharing, trust, shared awareness, and team performance dynamics—of different C2 approaches and organizational structures (CCRP, 2010, 2011). Since its initial development, ELICIT has been expanded upon with the addition of sensemaking virtual agents based on models of human information processing and collaboration behaviors (Ruddy, Wynn, & McEver, 2009; Smart, Richardson, Sycara, & Tang, 2014). Besides acting as a testbed for studying models of human and team cognition, the agent-based capabilities of ELICIT have allowed for large-scale experiments to validate much of the theoretical and human-in-the-loop experimental work on C2 Agility (Manso, 2012).
Furthermore, the ELICIT platform has been used to study, for instance, how organizations adapt to errors and delays in communication networks (Chan & Ivanic, 2010), how collective performance is affected by seeking and using information in the context of different organizational structures and communication patterns (Friedman, Bernstein, & Lazer, 2010), and as a pedagogical tool in civilian or military classrooms to highlight the strengths and weaknesses of hierarchical and edge organizations as they relate to problems that pose different challenges in terms of required levels of information-sharing and shared situational awareness (Ruddy, 2011).

2.4 Team Cognition and Performance

Teams— in organizational contexts rather than social, sports, or other settings—are temporary constellations of interdependent members with different roles, skills, and responsibilities who cooperate in order to achieve a common goal (Mathieu, Maynard, Rapp, & Gilson, 2008; Salas, Dickinson, Converse, & Tannenbaum, 1992). Teams differ from groups in that groups typically lack specialized roles or limited time frames. Although groups also work towards a common goal, special emphasis is put on social aspects such as shared beliefs, values, and norms regarding their domain (Berggren, 2016). The challenges faced by teams are essentially infinite in their variability, ranging from stable and predictable environments to complex and intense. This is reflected in the way team configurations can vary from homogeneous to heterogeneous depending on the team capacities required by a given task (Mathieu et al., 2008).

Whereas cognition refers to mental processes and capacities—such as perception, decision making, memory, etc.—in individuals, team cognition refers to cognitive activities within teams (Berggren, 2016). These activities—e.g. information management, communication, coordination, and cooperation—allow teams to perceive cues, make plans while drawing on prior knowledge and experience, and design and implement solutions to problems and tasks (Berggren, 2016; Cooke, Salas, Kiekel, & Bell, 2004).

Team cognition is typically described from two different perspectives. First, from the shared cognition perspective, team cognition is viewed as the approximate aggregate sum, mean, or similarity of the cognitive processes, knowledge representations, and constructs of its constituents (Berggren, 2016; Cooke, Gorman, Duran, & Taylor,
In this sense, team cognition is a product (Fiore & Salas, 2004). One such product could be a shared mental model. A mental model is an individual’s organized knowledge structure (containing domain knowledge, understanding, experience, etc.). A shared mental model refers to the overlap or similarity between converging mental models within the team (Berggren, 2016; Cooke et al., 2007). This is an example of what Cooke et al. (2004) refer to as a ‘collective’ view of team cognition. Shared cognition is based on Input-Process-Output (IPO)-models in which input variables (e.g. individual characteristics and competencies, task and organizational structure, or environmental contexts) affect team processes (goal-oriented team interactions), the study of which can explain the outputs (e.g. team performance outcomes or affective reactions such as satisfaction or commitment) (Berggren, 2016; Mathieu et al., 2008).

The second perspective on team cognition is interactive team cognition (Cooke, Gorman, Myers, & Duran, 2013). In this view, team cognition is not information processing in terms of encoding, storing, or retrieving information (as the aggregate cognitive activity of the team members) as argued by the shared cognition perspective. Instead it focuses on the interactions between team members - their communication and coordination (Berggren, 2016; Cooke et al., 2013). Team cognition is therefore believed to be an interactive process—an emergent phenomena—rather than a product or property of the team (Cooke et al., 2013). Although shared cognition also takes processes into account, these are considered to be mediators between input and output. Also, shared cognition is measured by querying team members at each of the I-P-O stages (Berggren, 2016). In contrast, interactive team cognition is not concerned with the input or output; the interest lies solely with the qualitative aspects of the processes that drive the interactions and dependencies between the inter-individual and intra-individual levels of the team (Berggren, 2016; Salas, Rosen, Burke, Nicholson, & Howse, 2007). This represents what Cooke et al. (2004) calls a 'holistic' view of team cognition.

Understanding team cognition (shared or interactive) is important because it is strongly linked with team effectiveness and performance (Cooke et al., 2007, 2013, 2004; Mathieu et al., 2008; Salas, Cooke, & Rosen, 2008; Salas et al., 2007). Specifically, with training and experience, teams become more coherent and develop more stable and uniform mental models of their operational environments and tasks.
This, in turn, enable teams to achieve higher levels of operational effectiveness and performance (Cooke et al., 2007, 2004). Effective communication—as a function of team cognitive processes—is a key factor in high-level team performance (MacMillan, Entin, & Serfaty, 2004). To the extent that the overhead costs of communication can be minimized (e.g. reducing its necessity or its resource demand in terms of time requirements and cognitive strain), team performance can be enhanced (MacMillan et al., 2004). Besides training to improve team cognitive capacities, team structure (i.e. C2 approach) also affects communication requirements, and therefore team performance (MacMillan et al., 2004).

2.5 Background synopsis

C2 has during its previous years mainly functioned through the deconstruction of large complicated endeavors into smaller, less complicated subtasks which can be accomplished by specialized task groups. To accommodate this approach to problem solving, C2 organizations have adopted the use of deconflicted hierarchies to ensure that task groups are focused solely on their task, thus completing the task faster and so also achieving operational success faster. Although slight variations may occur in regard to exact composition and functions included in the organization, this deconflicted approach to C2 has been the standard overall.

Heralded by the 21st century, the Information Age has drastically changed the operational environments facing C2 systems. The endeavors associated with the modern world are rife with uncertainty and complexities which render the deconflicted hierarchical approaches of the past ineffective. It has therefore become a main focus of C2 practitioners and researchers to develop methods to help C2 systems cope with, and operate effectively under, complex conditions. One of the capabilities identified as central to a C2 system’s success when operating in complex endeavors is the ability to cope with, adapt to, and beneficially exploit changes in circumstance. This ability has come to be termed C2 Agility.

C2 Agility builds on two other theoretical C2 concepts; the C2 Approach Space, and the Endeavor Space. The C2 Approach Space refers to a model containing all possible C2 approach constellations as defined according to three dimensions; Allocation of decision rights, Patterns of Interaction, and Distribution of Information. C2
Approaches are here placed along a diagonal line stretching from deconflicted organizations in one bottom corner, to edge organizations in the opposite top corner. The Endeavor Space serves a similar modeling purpose to the Approach Space but instead aims to encapsulate all possible endeavors a C2 system may contend with. The Endeavor Space argues that endeavors may differ along three dimensions; Tractability, Dynamics, and Dependencies. C2 Agility theory postulates that for an entity to be considered agile it must be able to adapt its C2 approach to suit the requirements set by the current operational circumstance. To be agile therefore requires knowledge of which C2 approach is best suited for which operational circumstance. That is, it requires the mapping of each of the C2 approach in the C2 Approach Space onto a specific region of the Endeavor Space.

Conducting research using simulations is a long standing and verified practice as it is associated with a number of benefits, such as allowing experimentation on situations which would have been otherwise inaccessible due to being either too dangerous or expensive to set up. Simulation refers to the process of creating a model of a real-world object or phenomenon (simuland) based on available knowledge of that object (referent), a model which may then be executed in a simulator to generate a simulation of the target object. ELICIT is a simulator intended to be used for studying team performance and behavior dynamics in different C2 approaches. ELICIT allows researchers to organize participants in various C2 constellations (e.g. a deconflicted hierarchy) and study the effects the chosen C2 approach has on various team cognition constructs, e.g. shared situational awareness, and the effects the teams structure has on team performance. The affordance allowed by the team’s structure has an immediate effect on the forms of interactions that may take place between teammates, impacting the conditions under which team cognition may emerge and affect team performance.
3 Method

To test whether the Endeavor Space dimensions could be instantiated and studied in ELICIT we first needed a tool designed to capture how these dimensions are perceived by people and organizations. As no such tool yet exists we first had to develop one. In this method section we explain the design, development, and validation of a functional prototype questionnaire—the Endeavor Space Subjective Assessment Instrument (ESSAI)—that measure the three Endeavor Space dimensions (tractability, dynamics, and dependencies) as experienced by responders. We will also present the details of a human-in-the-loop experiment where human participant teams completed three scenarios in the ELICIT simulator and reported on their experiences using the ESSAI questionnaire.

3.1 The Endeavor Space Subjective Assessment Instrument

We will now introduce the ESSAI tool by describing its development and validation process, its scoring system, and its application.

3.1.1 Design rationale

The Endeavor Space—as currently formulated—is intended as a conceptual space that encapsulates all possible mission circumstances that a C2 system may be faced with. Although the endeavor space relies on objective dimensions (Coupling, Tractability, and Dynamics) to capture the characteristics of different endeavors, the interpretation of where an endeavor is positioned within the endeavor space is—as of yet—not an objective assessment. To appropriately manifest C2 agility an entity must make an assessment regarding an endeavors position within the Endeavor Space and the suitability of the currently deployed C2 approach. Therefore, the question of how the Endeavor Space is perceived is to a certain degree a matter of subjective interpretation on part of the entity. An entity perceives the current circumstance of the endeavor and forms a hypothesis regarding the endeavor’s current position within the Endeavor Space, based on this assessment the entity implements the C2 Approach best suited for managing that particular set of circumstances according to the C2 Approach Space.

The design rationale behind the ESSAI tool (see Appendix A) is thus that a subjective measurement is required in order to assess the implementation of the Endeavor Space
dimensions in the ELICIT scenarios. ELICIT act as the problem domain that provides the circumstances from which participants will draw conclusions regarding the endeavor’s position in the Endeavor Space. If the questionnaire items in the ESSAI tool accurately capture the dimensions, it should capture the participants’ subjective experience of the Endeavor Space dimensions. Thus, if the participants subjective assessment of the Endeavor Space aligns with the configurations made in ELICIT, then the argument could be made that ELICIT is successfully able to simulate the Endeavor Space. Figure 16 below illustrates the intended process where the problem (ELICIT) is observed, a conclusion is drawn regarding the endeavor’s position along the three dimensions (the Endeavor Space), this assessment is captured by the ESSAI.

![Diagram](image)

Figure 16. The operator, having interacted with ELICIT, makes a mental assessment of the complexity of their endeavor, and reports this assessment using the ESSAI.

At the time of writing, the Endeavor Space is still in the early stages of its development. The definition of its dimensions (Tractability, Coupling, Dynamics) and their composition are topics of continuous debate. Therefore, the first step in the development of the ESSAI tool was to arrive at definitions of the dimensions that were adequate for the purpose of developing questionnaire items that were relevant and able to capture the aspects of the dimensions that could be subjectively
perceived. We will briefly account for how each of the Endeavor Space dimensions were interpreted and implemented to create the ESSAI tool.

One of the bigger issues we faced had to do with the Complexity/Tractability dimension. It was the opinion of the authors that using the term complexity for this dimension was confusing as both the coupling of components and the dynamic characteristic of a system are components of complexity (see e.g. Brehmer and Dörner (1993), and Braarud and Kirwan (2011)). The intended use of the Endeavor Space as a blueprint for assessing which C2 Approaches are most suited for dealing with certain circumstances is therefore—in part—a matter of understanding which approach is best suited for dealing with a certain level of complexity. Thus, there is an element of recursiveness when using complexity as a dimension; An endeavor’s overall complexity becomes a function of how its complexity is rated along the complexity dimension. Tractability can be viewed as a component of complexity where, depending on how easy it is to grasp which components constitute a problem, the easier it is to understand the problem, making the problem less complex. Conversely, if it is difficult to grasp, the problem becomes more difficult and complex. Tractability viewed as a component of complexity thus felt more appropriate than viewing complexity as being a component of complexity itself. The questionnaire items for the tractability dimension were therefore aimed at gauging the participants’ perception regarding their understanding of the problem, and if they felt they could make inferences regarding the state of the problem based of their understanding.

The contributions of dynamics to the overall complexity of a problem are expressed in the endeavor’s propensity to change, the magnitude (i.e. the severity and/or reach of the effects) of those changes, and the time pressures associated with acting within a dynamic environment. The questionnaire items related to dynamics were therefore designed to gauge the subjective assessment of these three experiential aspects. That is, to which degree did the participants feel they had to change their plan due to changes in the system. When a change occurred, did the participants perceive the change and the consequences it had for their planning as drastic, and to what degree was the effects of the change perceived as surmountable.

Coupling/causality (dependencies) is interpreted as the degree to which components of the problem required the actors to interact with each other in order to complete
their tasks. A tightly coupled problem would increase the need for interaction between actors as each component becomes more dependent on the others, thus increasing the need for interaction to solve the problem. Another effect of increased coupling would be that tasks would become more sequential as their relationships became more casual. The level of coupling would also affect the degree to which the choices and actions of others in the organization affected each individual actor. A loosely coupled problem would therefore be more modular where each problem can be approached independently, reducing the need for interaction between actors and promote more independent work as the actions of individual actors most likely will not affect the overall organization. The questionnaire items were therefore focused mainly on gauging the extent to which participants felt a need to communicate with their teammates to complete the tasks, experienced the tasks as being sequential in nature, and to what degree the participants felt they were affected by the actions of others.

When formulating the questionnaire items a decision also had to be made regarding at which level of description the items would be formulated.

A local approach could be used, where items are formulated in a domain specific manner aiming to capture the dimensions as experienced specifically in through tasks in ELICIT. Or, the second option would be to use a global approach where the items are broadly formulated, aiming to capture the experience of the dimensions regardless of the task domain. As covered in the end of the Endeavor Space section (see section 2.2.1), the dimensions are by necessity defined in a global fashion in order to be applicable to as many endeavors as possible. Therefore, it was decided that the ESSAI tool should follow the same principle. To allow the tool to be applicable in a variety of operational environments the questionnaire items must be focused on generic task characteristics which can be argued to evoke subjective experience of the Endeavor Space dimensions. In similar fields where task content cannot be deterministically connected to the theoretical concept that is to be measured, it is recommended to use questionnaire items that are focused on generic task characteristics rather than specific task elements (Salmon et al., 2008).
3.1.2 Structure and scoring rationale

The ESSAI tool consists of 14 question items. Three items addressing Tractability, four items addressing Dynamics, and three items addressing Dependencies. The last item, isolated from the dimensions, asks respondents to rate their perception of task difficulty. Each dimension is connected to a control question. The control questions are formulated so that it is the opposite valence of the dimension items. For instance, the control question for the dynamic dimension would be “Did you perceive the task as static?”. The control questions are used to ensure that each dimension is understood correctly. This may be analyzed by performing a correlation between the dimension score of interest and its control question. A negative correlational relationship would indicate that the dimension has been interpreted and understood as intended.

Each item was formulated as a statement (see the example in Figure 17). Respondents report their level of (dis)agreement using a 7-point Likert scale ranging from Entirely Disagree (scoring 1) orFully Agree (scoring 7).

![Figure 17. Example of questionnaire item from the ESSAI tool.](image)

The total score for each dimension is calculated through aggregating each questionnaire item for that particular dimension. For instance, the total tractability score is the aggregate mean of items 1-3. All dimensions are therefore scored between 1 and 7. Furthermore, computing a total complexity score is a matter of aggregating the mean total of the three dimensions into one overall complexity score.

However, it is worth noting that in terms of understanding which position an endeavor holds in the Endeavor Space, one should refrain from using an aggregate ESSAI score. As all the dimensions are a part of complexity, an aggregate score would describe only overarching differences in complexity between endeavors but fail to explain why or how two endeavors may be differently positioned in the Endeavor Space. Thus, for the purposes of mapping endeavors to the Endeavor Space, the overall complexity of each
endeavor is of secondary interest - the primary interest being how the two can be described or rated on each complexity dimension.

3.1.3 Testing and iteration

Upon completing the initial draft of the ESSAI tool it was decided that a preliminary testing phase was needed to ensure that the questionnaire items were well formulated and understandable. Since the purpose of this initial testing phase was only to gather feedback on the structure and wording of the ESSAI tool, the initial testing phase only included five participant testers and was conducted in-house with employees at FOI.

The reason behind conducting the initial testing in-house at FOI was due to FOI employees extensive experience of working with and developing research tools. We therefore considered them subject matter experts. Another reason for conducting the test in-house was due to FOI employees having some familiarity with system theory and related theoretical concepts like dynamics, complexity, and dependencies, which would allow them to provide feedback on use of the terms in ESSAI.

To test the ESSAI tool the participants were tasked with judging the three Endeavor Space dimensions (tractability, dynamics, dependencies) based on illustrations of complex dynamic systems. However, one hindrance associated with this approach was that the questionnaire items for the initial draft were formulated in a way that the questions only made sense if the responder answering the questions had participated in an activity (e.g. an ELICIT trial). Since the five participants would only be observing the illustrations rather than interact with them, certain items had to be adapted to fit this purpose. Due to these changes and the small sample size of the initial test, the participants’ ESSAI scores themselves were of little interest or use and will therefore not be reviewed here.

Three illustrations of complex dynamic systems were used during the test. Each illustration was taken from complexity-explorables.org, a website dedicated to hosting illustrations of complex dynamic systems. The three chosen illustrations were called Yo, Kohonen, Into the Dark, and Critically Inflammatory. These were chosen based on their perceived goodness-of-fit for the study’s purpose, that is whether or not it would be possible to hypothesize about the complexity of the systems by only viewing the illustration. Each system consisted of several interacting elements or agents
reacting in different ways to some form of stimuli. Due to this the system illustrations were also chosen based on their perceived difference in dependency, dynamics, and tractability so that for each illustration the participants would be viewing a system that behaved arguably different from the other systems that were shown.

Each participant viewed the systems one at the time for about 5 minutes or until the participant felt that they had an understanding of the system’s behavior. The ESSAI tool was then administered, the participant submitted their answers, and then proceeded to the next system illustration. This process was repeated for all three illustrations. Upon completing the third system assessment, the participants were briefly interviewed regarding their thoughts on the ESSAI as a tool for evaluating the perceived system complexity.

3.1.3.1 Results and revisions

As the focus of this initial testing was solely to gather feedback on the formulation of the questionnaire items and the questionnaires structure, no attention was paid to the participants’ scoring of the dimensions. How the participants scored each illustration was also of second interest as the domain it was used on—evaluating visual illustrations—was so different from the intended domain (see, ELICIT section 2.4.2). However, the feedback gathered on the ESSAI was of utmost interest. Primarily it was learned that the use of certain contextual and abstract words or phrases—e.g. “system”, “develop over time”, etc.—in the formulation of certain items were confusing and open to interpretation. It was thus decided that the tool had to be looked over once more to reformulate certain items and reduce the amount of words that required contextual knowledge, or, wherever possible, provide the context needed to understand the question. Certain feedback on negative aspects of the tool was deemed to be of secondary interest as that feedback was most likely due to a mismatch between the domain in which the tool was tested (an observation task) and the domain for which the questions were developed (a participation task). It was argued that issues related to this would fix themselves once the tool was used in a task in which the participants would be participating in an activity before being administered the questionnaire.
3.1.4 Validation

Before conducting the ELICIT experiments it was decided that a validation test phase was needed to test the construct validity of the revised ESSAI tool, i.e. that its aggregated items are measures of perceived Endeavor Space dimensions. Also, the test was necessary to ensure that the ESSAI tool works as intended in a participatory setting. For these purposes, a small experimental study was conducted. The board game Pandemic, designed by Matt Leacock and published by Z-Man Games in 2008, was chosen as a testbed.

3.1.4.1 Pandemic

Pandemic is a turn-based cooperative board game where two to four players are tasked with curing four diseases before the world ends. Each player is given one out of seven available roles—e.g. medic, quarantine specialist, coordinator, etc.—each with their own special abilities. On their turn, players perform four actions, such as travelling in different ways, curing infections, give away or take city cards, or curing diseases. They then draw two player cards. These cards can be either city cards, epidemic cards, or special action cards. City cards are color-coded to each of the four diseases: black, blue, red, or yellow. At the end of their turn, players also draw two to four infection cards, the number increasing each time an epidemic card is drawn from the player card pile. Like the city cards, the infection cards indicate which city should be infected by a particular disease. A city can only hold three infection cubes of a disease at any given time. Each infection cube can be treated and removed by expending a player action. Should a city holding three cubes receive a fourth cube of the same color, an outbreak occurs where all adjacent cities get one cube each while the initial city remain at three. A fail condition in the game is that eight such outbreaks occur. A second fail condition is that the pile of player cards is depleted, and the third fail condition is that all available infection cubes of any disease are placed on the map. There is only one way to win: to cure all four diseases. This is done by collecting five colored city cards—e.g. red, corresponding to the disease—travelling to a research station (which players can build across the world) and, as a player action, spend those five city cards to find a cure for that particular disease. Once all four diseases have been cured in this way, the team wins.
The game can be played on three levels of difficulty by deciding how many epidemic cards are in the player deck. When an epidemic card is drawn, the bottom card in the infection card pile is drawn and three infection cubes are placed on the corresponding city. Next, the newly drawn infection card and the rest of the infection card discard pile is mixed thoroughly and placed on top of the infection card pile. Finally, the player ends his or her turn as normal by drawing two to four cards from the infection card deck and infect the drawn cities with one infection cube. The obvious danger here is that previously infected cities can be further infected, including cities that already have three infection cubes. The threat of widespread chain-reacting outbreaks loom and player teams must be prepared to adapt quickly to defuse any potentially dangerous situations. Naturally, teamwork and cooperative planning is vitally important if the team is to have any success.

3.1.4.2 Participants

Sixteen participants (9 men, 7 women, mean age = 24.6 years, SD = 1.55) were included in the study, and were divided into four teams of four. All participants were university students. Some of them were familiar with the Pandemic board game but none reported having extensive previous experience of playing it. No monetary compensation was offered to the participants.

3.1.4.3 Ethics

The study was conducted in accordance with the Swedish Research Council’s guidelines regarding ethical research (Vetenskapsrådet, 2002).

All participants were required to provide their informed written consent before participating in the study. The consent form (see Appendix B, in Swedish) briefed the participants on the purpose and contents of the study, and that their data would be anonymized and kept confidential to any outsiders. It further informed them that their data was only to be used for academic purposes. It also informed them that participation was entirely voluntary, and that they could rescind their participation at any time and for any reason, in which case their data would be destroyed.

3.1.4.4 Scenarios

To test the ESSAI tool, two Pandemic scenarios were designed - one Easy, and one Hard. The scenario order was counterbalanced; Two groups played the game starting
with the easy scenario, ending with the hard scenario, and the other two groups started with the hard scenario and ended with the easy scenario. To achieve an appropriate difference between the scenarios, a number of manipulations to the game rules and gameplay mechanics were made. Changes to the game were primarily done with the intent to manipulate the Endeavour Space dimensions so that each scenario would differ along the three dimensions. The Dynamics dimension was manipulated by varying the number of epidemic cards in the deck. A higher number of epidemic cards would introduce more variability and changes into the game as well as increase time pressure through faster escalation and increased risk of outbreaks. Tractability was manipulated by decreasing the number of game components involved in the game. Dependencies were manipulated through changing the game roles allowed for each scenario; The roles available to the team were argued to impact the strategies available to the team in terms of cooperation and planning. Each scenario and their respective manipulations will be summarized below.

Easy scenario set-up:

- Four epidemic cards
- Roles: Medic, Quarantine specialist, Engineer, and Coordinator
- No Event cards

In the easy scenario, the minimum number (four) of required epidemic cards were used. Since epidemic cards is the direct control of difficulty in Pandemic, they control the spread rate of disease. As a team, you are more likely to fail the game and lose more quickly if more epidemic cards are present in the deck. Furthermore, the number of epidemic cards also represent the Dynamics dimension in that fewer epidemic cards means that there is a reduced rate of change and magnitude of change. Roles were chosen based on their impact on the Dependencies dimension and their usefulness in Pandemic. The only change from the Hard scenario is that the Medic role is allowed, which makes the game significantly easier. The Medic role is also largely an independent actor and thus reduces the need to coordinate with teammates to cure disease, resulting in lower Dependencies. Event cards were removed because they constitute another point of consideration. Granted, they make the game easier in the sense that they afford the team certain abilities that are helpful, but from a complexity
standpoint they increase the number of components to keep track of and plan for. Removing them would—in our view—increase Tractability.

Hard Scenario set-up:

- Six epidemic cards
- Roles: Scientist, Quarantine specialist, Engineer, and Coordinator
- All five Event cards are included

In the Hard scenario, all six epidemic cards were present. Not only did this make the game innately more difficult, but it also increased the dynamic rate of the game since larger changes occurred more frequently (more epidemic cards were drawn) and small changes (city infections) became more frequent in the sense that (potentially) more cities were infected each turn. The Scientist role was added both to remove the powerful Medic but also because her ability (fewer cards required to cure disease) provided a different challenge to the team in terms of coordinating their actions. From personal experience the authors also know that her ability may be a distracting factor as teams tend to over focus on transferring cards to her rather than curing diseases themselves. This may cause a lot of time and effort to be wasted. For the Hard scenario Event cards were re-introduced to the deck to increase the number of components to consider (even though they are technically helpful to the team), decreasing Tractability.

In both scenarios, the teams were given 30 minutes to cure all four diseases. A set playtime was decided to ensure that each team would receive the same amount of time for each scenario. It is worth noting also that, in certain cases, 30 minutes is not enough time to successfully complete the game. It was argued however that successfully completing the game was not needed for the participants to experience the game and get a sense for how the game’s different components worked and interacted with each other.

3.1.4.5 Design

The validation study used a within-group design. The independent variable was the scenario used (Easy, Hard). The dependent variables were the ESSAI response data, specifically the calculated scores for Tractability, Dynamics, and Dependencies, along with ratings of subjective difficulty.
3.1.4.6 Materials

In terms of materials only pen and paper, the ESSAI questionnaire, and the Pandemic board game were used.

3.1.4.7 Procedure

Participants were welcomed and thanked for wanting to participate in the study. Before introducing participants to Pandemic, they were asked to read and sign a consent form containing a brief summary of the study’s purpose and information regarding their rights as participants.

Once all participants had filled out the consent form, they were verbally introduced to the Pandemic board game. The rules, goals, and functions of all set pieces were explained. All teams received the same introduction. No hints were given to the difficulty of each scenario. The teams were encouraged to ask questions during the introduction if clarifications were needed. Once the team felt that they had no more questions and seemed prepared they began playing their first scenario.

Each scenario had a designated playtime of 30 minutes, success or not the scenario was ended after 30 minutes had passed. At the 15-minute mark for the Easy scenario, and at the ten-minute mark for the Hard scenario, the game was paused, and the participants were asked to answer the first questionnaire. Once all participants had indicated they were finished with the questionnaire the game session was resumed.

Once the scenario was finished the participants were once again asked to answer the rest of the questionnaire. The teams were allowed a short respite for ten minutes between scenarios as the experiment leaders had to reset the board and reconfigure the difficulty settings. The teams were encouraged to discuss strategy in between scenarios. Once the board had been reset the participants were told that they could start playing again, and the same procedure regarding the questionnaires was followed for the second scenario.

Once both scenarios were complete, and all questionnaires had been answered, the participants were once again thanked for their participation and then dismissed.
3.1.4.8 Results

Prior to conducting any statistical analyses, the data was screened for missing values or other irregularities. This process resulted in that one data point was excluded because of missing data values. No tests were conducted to detect or remove outliers.

First, the overall subjective difficulty of the two scenarios was calculated. The Easy scenario received a mean difficulty score of 5.3 ($SD = .82$), and the mean difficulty score of the Hard scenario was 5.47 ($SD = .72$), as rated by the participants. This indicates that the Easy scenario was indeed perceived as easier compared to the Hard scenario. Bear in mind, however, that the sample size of this validation study was limited.

Next, analyses were conducted to investigate whether the aggregate mean of the questionnaire items related to each Endeavor Space dimension correlated with its respective control question. Again, assuming that the questions were correctly interpreted, one would expect to see negative correlations between the dimension scores and their control questions, since the control questions were formulated to reflect the opposite valence of the dimension questions. The Pearson's $r$ correlation matrix in Table 1 shows the relationships between these variables, along with the subjective difficulty variable.
Table 1. Correlation matrix showing the relationship between variables. The correlation value (Pearson’s r) ranges from -1 (negative correlation), to 0 (no correlation), to 1 (positive correlation). Effect sizes can be interpreted as ±.1 = weak correlation, ±.3 = moderate correlation, and ±.5 = strong correlation (Field, 2009).

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Tractability</td>
<td>.16</td>
<td>-1.0</td>
<td>-.41*</td>
<td>.06</td>
<td>.19</td>
<td>-.19</td>
</tr>
<tr>
<td>2. Dynamics</td>
<td>.39*</td>
<td>-.18</td>
<td>-.56**</td>
<td>-.07</td>
<td>.39*</td>
<td></td>
</tr>
<tr>
<td>3. Dependencies</td>
<td>-.02</td>
<td>-.44*</td>
<td>-.38*</td>
<td>.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Tractability Control</td>
<td>.33</td>
<td>-.16</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Dynamics Control</td>
<td>.14</td>
<td>-.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Dependencies Control</td>
<td></td>
<td>-.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * correlations are significant at the p = .05 level. ** correlations are significant at the p = .01 level.

As can be seen in the matrix in Table 1 all three aggregate dimension scores are negatively correlated with their respective control questions: Significant correlations were found between Tractability and its control question, \( r(30) = -.41, p = .021 \); Dynamics and its control question, \( r(30) = -.56, p = .001 \); and Dependencies and its control question, \( r(30) = -.38, p = .038 \). These results indicate that the questions related to each dimension were interpreted as intended.

Interestingly, further significant correlations were found between Dynamics and Dependencies, \( r(30) = .39, p = .032 \). This result could be an artefact of the limited sample size but would not be entirely unexpected due to the overlapping nature of the two constructs. For instance, events and changes may have further-reaching consequences in a tightly coupled system, which can result in an increased sense of system dynamics. This relationship would also explain the significant correlation between Dependencies and the Dynamics control question, \( r(30) = -.44, p = .012 \).

Finally, the correlation matrix in Table 1 shows how each dimension relates to subjective difficulty. Tractability showed a non-significant and weak correlation with
subjective difficulty, $r(30) = -0.19$, $p = 0.314$. A statistically significant and moderately sized correlation was found between Dynamics and subjective difficulty, $r(30) = 0.39$, $p = 0.029$. Lastly, a non-significant yet moderately strong correlation was found between Dependencies and subjective difficulty, $r(30) = 0.35$, $p = 0.054$. These results are somewhat expected as it was not entirely obvious how to reliably manipulate Tractability in Pandemic. Rather, it was believed that the perceived effects of the hypothetical manipulations of Dynamics—i.e. more epidemic cards, etc.—would be the most salient. The manipulations done to the Dependencies dimension, i.e. swapping the Medic role for the Scientist, seem to have had the desired effect in terms of perceived difficulty.

Based on these results, we believed that the ESSAI tool was adequately designed and reliable for use in the main ELICIT experiment, which is the topic of the next section.

3.2 ELICIT Experiment

The second phase of the study involved running participant trials in ELICIT and having the participants assess the complexity of three ELICIT scenarios using the ESSAI tool. This section will describe the ELICIT experiment trials themselves, the results of which are presented in the Results section.

3.2.1 Participants

Fifty-six students (35 men, 21 women, mean age = 24.4 years, $SD = 3.51$) were recruited to participate in the study. The only requirement for being eligible for participation was to be a university student. Participants were recruited from and via the personal social networks and contacts of the authors, and by personally approaching random student groups on campus. The 56 included participants were subsequently divided into eight teams where team affiliation was decided primarily by the simultaneous availability of the required number of participants. Teams were also formed as initial participants in turn recruited their friends and classmates. No participants reported having any previous experience of the ELICIT simulator. Participation was rewarded with a movie ticket. Biscuits were also served during sessions.
3.2.2 Ethics

Like the ESSAI validation study, the ELICIT study was designed to comply with the Swedish Research Council’s guidelines for ethical research (Vetenskapsrådet, 2002). Upon their arrival, participants were asked to read and sign a consent form (see Appendix C, in Swedish). This consent form informed the participants about the nature and purpose of the study. It further informed them that participation was entirely voluntary and anonymous, and that there were free to withdraw their participation at any time - in which case their data would be destroyed. Participants were also informed that their data would be anonymized and kept safe from third-party entities, and that their data were only to be used for academic purposes.

3.2.3 Design

A within-group design approach was applied where each team encountered all three scenarios. The order in which team encountered the scenarios was counterbalanced to control for learning effects. Four teams completed the three scenarios—Easy (E), Medium (M), and Hard (H)—in the order E-M-H, three teams in the order M-H-E, and the last team completed the scenarios in the order H-E-M.

The independent variable was the ELICIT scenario (three levels: Easy, Medium, Hard). The dependent variables used were the team-level aggregate ESSAI response data—measures of Tractability, Dynamics, and Dependencies, plus subjective difficulty—and the aggregated team performance scores as extracted from ELICIT log data.

3.2.4 Instruments

The ESSAI tool (described in section 3.1) was used to measure how the participants perceived the complexity of the scenarios. Additionally, NASA-TLX and SART questionnaires were administered.

NASA-TLX (see Appendix D) is a post-trail subjective workload measurement developed by Hart and Staveland (1988). The NASA-TLX calculates workload as an overall score that itself is taken from the assessment of six subscales (mental demand, physical demand, temporal demand, performance, effort, frustration) that are measured on a 100-point scale with 5-point increments. These subscales are scored that are calculated into an overall workload score. The NASA-TLX can be used in
multiple different ways as the previously mentioned subscales can be adjusted to fit the
task that is being assessed, e.g. the physical demand subscale can be removed from the
questionnaire if it is not relevant to the task, this can be done without affecting the
validity of the measurement. It is then referred to as Raw-TLX (Hart, 2006). For the
present study the Raw-TLX version was used as the physical demand subscale was
removed since it was not relevant for the ELICIT task.

SART (see Appendix E) is a post-trail subjective situational awareness measurement
developed by Taylor (1990). The questionnaire measures situational awareness as a
sum of ten factors; Instability of situation, complexity of situation, variability of
situation, arousal, concentration of attention, division of attention, spare mental
capacity, information quality, information quantity, familiarity with situation. Each
factor is assessed on a 7-point Likert scale (1 = low, 7 = high). These factors are
aggregated into a total sum which represents the participants situational awareness. It
should be noted that the SART questionnaire used for this study only included nine of
the ten above mentioned factors, as presented in Salmon et al. (2008).

Although the NASA-TLX and SART questionnaires were used in this study, analyses of
their results are not included here.

3.2.5 Equipment and materials

Eight laptops were used. One was used as the central server, and one for each of the
seven members of a team. These were all connected to a LAN via a switch and network
cables. USB mice were connected to the player computers. The server laptop ran Oracle
Glassfish Server v.3.1 to host the ELICIT v.2.6.2 software. Microsoft Internet Explorer
11 was used to access the ELICIT graphical user interface on both the server and client
computers. The ELICIT Log Analyzer was used to access and explore the log files.

Consent forms and ESSAI questionnaires were printed on paper and signed with
pencils.

A regular smartphone was used to keep the time of the 10+5 minute scenarios.
3.2.6 Scenarios

The teams were arranged in a classical hierarchical organizational structure where the top manager was in charge of two middle managers, each of whom in turn had two subordinates. As described in the ELICIT section, the goal of the organization(s) is to figure out the *Who, What, Where, and When* of an impending attack. Each bottom-level member was tasked with solving one of these four problems. Each middle manager had to solve both of the problems given to his or her two subordinates, and the top manager had the overarching responsibility of finding the solutions to all four problems.

Three scenarios were used; Easy, Medium, and Hard. These were created using the various configuration settings available in ELICIT. The configurable settings were conceptually mapped to each of the Endeavor Space dimensions. These were the following:

- **Tractability**
  - The total number of factoids in each scenario.
  - The signal-to-noise ratio, i.e. the amount of useless or distracting information compared to the amount of relevant and important pieces of information in the factoid set.

- **Dynamics**
  - Time available (scenario length).
  - Number of factoid distribution waves.
  - Frequency of factoid distribution waves.

- **Dependencies**
  - The amount of non-organic information, i.e. the amount of information being distributed to the “wrong” person, requiring more communication and cross-level information sharing.

These configuration settings represent the direct manipulations done to each dimension, but additional indirect manipulations were also applied. For instance, all of the settings directly associated with Dynamics were kept constant - the set time length for all scenarios was 15 minutes (based on internal tests), and all scenarios used three factoid distribution waves at five-minute intervals where the initial wave of
factoids was delivered at the very start of the scenario. Rather, the Dynamics differed between the scenarios as the result of the interactions between the base conditions determined by the direct Dynamics settings (i.e. 15 available minutes) and the stresses and requirements imposed by the Tractability and Dependencies dimensions. As the amount of information—and noise-to-signal ratio—increases (lower Tractability) along with the need for cross-level information sharing and communication (higher Dependencies), the Dynamics dimension is indirectly increased as the required amount of work (sorting, processing, sharing, etc.) per available minute increases, as the time limit is held constant. The three scenarios were constructed by adjusting the settings—directly or indirectly—in ways that would (in theory) evoke measurably different experiences related to each dimension.

The Easy scenario consisted of 34 factoids with a high signal-to-noise ratio (high Tractability). All factoids were delivered to the bottom-level workers, and each worker only received factoids related to their problem (organic information). Minimal horizontal communication was therefore needed. The information needed only to be sent vertically (low Dependencies). The high Tractability settings combined with the low Dependencies settings meant that the held-constant direct Dynamics settings—15 minutes, three factoid waves at five-minute intervals—provided ample time to sort, process, and share the available information, evoking a feeling of lower time pressure (low Dynamics) compared to the other two scenarios.

The Medium scenario also consisted of 34 factoids, but the signal-to-noise ratio was lower compared to the Easy scenario, meaning that more processing was needed to filter out the irrelevant information (medium Tractability). The factoids themselves were no longer solely distributed to the bottom-level workers - some were delivered to the middle managers. Furthermore, some non-organic information—i.e. factoids related to a different problem than your own—was introduced, requiring some cross-level sharing between different ends of the organization (medium Dependencies). Compared to the Easy scenario, the increased workload caused by the additional noise, combined with the increased need for cross-level sharing, resulted in more work needing to be done in the same amount of time, which should be experienced as an increased sense of time pressure (medium Dynamics).
Finally, the Hard scenario consisted of 68 pieces of information where the signal-to-noise ratio was even lower compared to the Medium scenario (low Tractability). The amount of non-organic information was also increased compared to the Medium scenario, as the factoids were distributed in such a way that substantial cross-level sharing was required to complete the tasks (high Dependencies). The increased information density combined with the increased need for communication and information sharing resulted in an increased amount of work needing to be done per available working minute compared to the Medium scenario. This would result in increased levels of stress due to the (indirectly) increased time pressure (high Dynamics).

Figure 18 illustrates the organizational hierarchy of the teams, and the tasks assigned to the members. It also shows—in concept—how the factoids related to each problem (in this case, the When problem) are distributed across the organization.

![Figure 18. Organization hierarchy and assigned tasks. The colored dots represent the factoids necessary to fully solve the When problem. Green dots represent the Easy scenario, orange squares represent the Medium scenario, and the red triangles represents the Hard scenario. Note that this is but a conceptual illustration of the distribution between the scenarios, rather than an accurate representation of the actual distributions.](image)
3.2.7 Scoring

Several score measurements and procedures were used. This subsection is dedicated to describing how these scores were elicited.

3.2.7.1 ESSAI rating scores

The ESSAI tool was administered twice in each scenario: mid-task, and post-task. The results of these two measurements were averaged to construct mean Endeavor Space dimension scores for each scenario. This approach of averaging ordinal—in this case Likert-type data—is a contested topic in the field of statistics. Some would argue that because the theoretical true distance between Likert items like 1 = “Definitely Disagree” and 2 = “Disagree” can’t be guaranteed to be equal to that of 3 = “Neutral” and 4 = “Agree”, mean values, such as 4.5, are meaningless (Grace-Martin, n.d.; Norman, 2010). However, this only affects the inferences one can draw about the underlying meaning of the Likert responses. It does not invalidate conclusions about means, differences, or any other qualities pertaining to the numbers themselves. Furthermore, even though Likert-type data tend to be—by its very nature—skewed, parametric statistical tests, including Pearson’s $r$ correlation coefficients and various $F$ tests, are quite robust to skewed or non-normally distributed data means (Norman, 2010). For our purposes, we were not interested in what the participants’ ESSAI responses meant, only whether differences in their responses—as a result of our Endeavor Space dimension manipulations—could be detected statistically. Therefore, it was our assessment that averaging individual participants’ ESSAI ratings, treating the construct measure as a continuous variable in standard parametric tests, was a useful and appropriate approach given the purpose of this study.

3.2.7.2 ELICIT performance scores

As explained in the previous section, the top manager was tasked with solving all four problems, each middle manager had to find the answers to two problems, and each bottom-level worker was assigned one of the four problems. Although participants were free to submit as many answers as they liked (including to problems assigned to others than themselves), only their latest answers (to their assigned problem or problems) were considered when calculating team scores. Each correct answer was rewarded with one (1) point for a maximum total of 12 points per scenario. The answer
to the “when” problem consisted of 4 components; the month, date, time, and time of
day (AM/PM). Therefore, each component was equal to a score of 0.25 points, a fully
correct answer would thus yield 1 point whilst partially correct answers would yield
0.25 points for each correct component. This approach was chosen as providing a
moderate score for partially correct answers would give a more accurate representation
of the teams’ performance than simply giving no points at all.

To determine if a player answer was correct or not, the answer key for each scenario
was used as a template to identify keywords which, if present in the player answer,
would count as correct. For example, while the answer key for the medium problem
below reference’s the “deltaland embassy” as answer to the what question, full points
would be rewarded to the participant if they had answered only *embassy* as it is the
one factoid that directly references the question of what will be attacked. However, no
deductions are made for answering “deltaland embassy” either. Below we will mark
each problem keyword for each scenario answer key.

Answer key for the easy scenario:

The **Violet group** (WHO) plans to attack a **financial institution** (WHAT) in
**Psiland** (WHERE) on **April 5 at 11:00 AM** (WHEN)

Answer key for the medium scenario:

The **Aqua group** plans to attack the **Deltaland embassy** in **Alphaland** on **May 27
at 3:00 PM**

Answer key for the hard scenario:

The **Magenta group** plans to attack the Southern **Oil Pipeline terminal** in
**Thetaland** on **August 15 at 11:00 PM**

Wrong answers would be therefore be considered answer which are considered
deviations from these keywords. For example, answering “oil rig” for the what question
in the hard scenario is incorrect as it references the wrong target, answering “pipeline
terminal” would have been correct. Answering “Magenta group or Aqua group” for the
hard scenario is also considered incorrect, although the answer contains the correct
keyword “Magenta group” it is deemed incorrect as it also includes “or Aqua group”.
The scoring system in the ELICIT log analyzer is tuned in a similar manner. However,
it is stricter towards free-text answers, e.g. “they will attack the deltaland embassy” would be marked as incorrect although being entirely correct. Therefore, the logs had to be processed manually to ensure that the scoring criteria set forth by the authors was followed.

3.2.8 Procedure

Participants were greeted upon their arrival and were asked to read and sign the consent form. Next, one of the authors verbally introduced the participants to the ELICIT platform. A basic script was used to ensure that all teams were given the same information. The introduction included an explanation of the hierarchy of the team and its capabilities and constraints. A whiteboard was used to illustrate this. The basic premise of the scenarios was explained; an attacker (e.g. the lion) would recruit a group (the who, e.g. the azure group) to attack a target (the what, e.g. a financial institution) in a certain country (the where, e.g. Psiland) at a given date and time (the when, e.g. April 5 11 AM). It was then explained how these problems related to the team’s hierarchy, and that teamwork and mutual sharing of information is key to achieving success. Next, participants were introduced to the ELICIT graphical user interface; the Inbox, Add to my Factoids, Share, Post, and Identify functions were explained, along with the functionality of the websites related to each problem. Participants were also informed that they would not be allowed to communicate orally while playing. Lastly, participants were told that they were free to ask any questions they might have. Once the team was ready, they began playing their first scenario.

When starting a scenario, a ten-minute timer was set on a smartphone. Once the ten minutes were up the authors paused the game (verbally, as sessions cannot be paused in ELICIT) and administered the questionnaire set containing the ESSAI, N-TLX, and SART instruments. Participants were instructed to fill out only the first ESSAI questionnaire (as the second, along with the other questionnaires, were to be done post-task). Once all team members were done filling out the first ESSAI, a five-minute timer was set, and the game was resumed. Countdown callouts were made at T minus three, two, and one minutes to alert participants as to give them time to submit their answers (if they had them). Once the five minutes were up the scenario was completed.
The participants were then instructed to fill out the rest of the questionnaire sets they had been given.

A short break (about ten minutes) followed once the first scenario was completed and all questionnaire had been filled out. Biscuits were served and teams could speak freely about whatever they liked, be it the scenario they had just completed, their tactics, or different matters entirely. Once ten minutes had passed or when the team felt ready to continue, the break was ended, and they completed their second and third scenarios in the same manner as the first with an approximated two-minute break in-between where a further round of biscuits was served.

Once the third scenario had been completed and all questionnaire sets had been collected the participants received their movie tickets and their due thanks for their willingness to participate. The session was then ended.

3.2.9 Analysis

All questionnaire data were manually entered into Microsoft Excel.

Team performance scores were extracted from the log files using the ELICIT Log Analyzer and submitted to Microsoft Excel.

Statistical analyses were performed in IBM SPSS Statistics 24, and an R-based open-source statistical platform called jamovi v.0.9.6 (see The jamovi project, 2019).
4 Results

In this chapter, we will first present the results of the statistical analyses conducted to assess the internal consistency of the ESSAI tool. The purpose of these tests was to determine whether the ESSAI tool was functioning as intended. This is a necessary step before presenting and analyzing the ELICIT performance and ESSAI assessment scores.

4.1 ESSAI internal consistency results

First, an exploratory factor analysis (EFA) was conducted to assess the structure and internal validity of the ESSAI tool. This procedure calculates whether and how question items group together to measure (load onto) latent variables (e.g. Endeavor Space dimensions). The ten main ESSAI items (questions 1-3, 5-8, and 10-12) were included in the model. Descriptive statistics of the observed variables are presented in Table 2, below.

Table 2. Descriptive statistics for observed variables included in the EFA. N = 335 (responses, three per participant). One case was excluded due to missing data.

<table>
<thead>
<tr>
<th>Question</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.72</td>
<td>5.67</td>
<td>4.72</td>
<td>3.55</td>
<td>2.91</td>
<td>4.96</td>
<td>4.47</td>
<td>5.37</td>
<td>3.25</td>
<td>6.01</td>
</tr>
<tr>
<td>SD</td>
<td>1.24</td>
<td>1.22</td>
<td>1.39</td>
<td>1.48</td>
<td>1.30</td>
<td>1.22</td>
<td>1.52</td>
<td>1.58</td>
<td>1.45</td>
<td>1.20</td>
</tr>
<tr>
<td>min - max</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
</tr>
</tbody>
</table>

The Keiser-Meyer-Olkin measure indicated a sufficient sampling size, KMO = .73 (‘good’ according to Field, 2009), however the individual KMO values of Question 1 and Question 2 were .41 and .38, respectively, which is below the acceptable limit of .5 (Field, 2009). Bartlett’s test of sphericity, $\chi^2 (45) = 901.63$, $p < .001$, indicated an adequate correlation structure between items. Three factors had eigenvalues greater than Keiser’s criterion of 1, and explained a cumulative 47.16% of the variance. These three factors were therefore retained. Based on the items that loaded onto the factors, the first factor was labelled Tractability, the second was labelled Dynamics, and the third was labelled Dependencies. Table 3 shows factor loadings after rotation.
Table 3. Summary of EFA results for the ESSAI tool. Factor loadings are post-rotation, representing regression coefficients.

<table>
<thead>
<tr>
<th>Item</th>
<th>Rotated Factor Loadings</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tractability</td>
<td>Dynamics</td>
<td>Dependencies</td>
<td></td>
</tr>
<tr>
<td>Question 1</td>
<td>.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 2</td>
<td>.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 3</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 5</td>
<td>.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 6</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 7</td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 8</td>
<td>.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 12</td>
<td></td>
<td></td>
<td></td>
<td>-.91</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>3.01</td>
<td>2.03</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>% of variance</td>
<td>25.92</td>
<td>15.55</td>
<td>5.69</td>
<td></td>
</tr>
</tbody>
</table>

Note. Extraction method: principal axis factoring; Rotation method: Oblimin with Kaiser Normalization; Factor loadings < .4 are suppressed.

Most of the items loaded onto the extracted factors as expected. A number of exceptions were found, however. Although questions 1-3 grouped together in the factor labelled Tractability, they were joined by Question 7 which did not load onto the Dynamics factor together with questions 5, 6, and 8. This suggests that Question 7 is more akin to the items designed to measure perceived tractability than those designed to measure perceived dynamics. Furthermore, questions 10 and 11 did not load onto any of the three extracted factors, including the factor labelled Dependencies which would represent their hypothesized latent variable. This result suggests that, besides Question 7, Question 10 and Question 11 need further refinement.

Furthermore, the results revealed weak correlations between the three latent factors. Tractability correlated with Dynamics and Dependencies by $r = .30$ and $r = .25$, respectively. Dynamics and Dependencies correlated by $r = .30$. To assess the influence
of these correlations, Table 4 presents the non-rotated factor loadings, which describes correlations between items and factors when not accounting for interaction effects.

**Table 4. Non-rotated EFA results for the ESSAI tool. Factor loadings are pre-rotation, representing correlation coefficients.**

<table>
<thead>
<tr>
<th>Item</th>
<th>Tractability</th>
<th>Dynamics</th>
<th>Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td>.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 2</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 3</td>
<td>.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 5</td>
<td>.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 6</td>
<td>.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 7</td>
<td>.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 8</td>
<td>.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 10</td>
<td>-.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question 12</td>
<td>-.87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. Extraction method: principal axis factoring; Rotation method: Oblimin with Kaiser Normalization; Factor loadings < .4 are suppressed.*

When the effects of other factors are ignored, factor loadings look much the same except that Question 10 now loads onto Dependencies. This result suggests that the intra-factor correlations affect how Question 10 (and by extension, Dependencies) is interpreted and rated. Question 7 still loads onto Tractability rather than Dynamics, and Question 11 is still unaffiliated to all three factors.

Overall, the results of the EFA suggest that the ESSAI question items—with the exception of Question 7, Question 10, and Question 11—belong together in their respective Endeavor Space dimensions, and that the dimensions are separate but related constructs as discussed in the Endeavor Space literature.

Next, reliability tests were conducted on each of the three Endeavor Space dimension constructs. All question items were kept in their respective intended dimensions. Question 7 was reverse scaled in the Dynamics dimension because of its positive
framing—and therefore mitigating effect—on Dynamics. The results are presented in Table 5, below.

Table 5. ESSAI dimension reliability matrix. Cronbach’s α values greater than about .7 indicate sufficient reliability (Field, 2009).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean (SD)</th>
<th>Cronbach’s α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tractability</td>
<td>5.37 (1.09)</td>
<td>.81</td>
</tr>
<tr>
<td>Dynamics</td>
<td>3.49 (.96)</td>
<td>.64</td>
</tr>
<tr>
<td>Dependencies</td>
<td>4.88 (.91)</td>
<td>.28</td>
</tr>
</tbody>
</table>

The individual contributions of the question items were calculated for each dimension. This revealed that dropping Question 3 from the Tractability dimension would increase the α to .84. For the Dynamics dimension, dropping Question 7 and Question 8 would increase α to .68 and .65, respectively, indicating that these question items need more work. Dropping Question 11 from the Dependencies dimension would increase the α of Dependencies to .54, suggesting there is an issue with the item. Overall, the Tractability factor is adequately reliable, but the Dynamics and (in particular) the Dependencies factors need further refinement.

Finally, Pearson’s r correlations were analyzed to test the Endeavor Space dimension constructs’ relationships to their respective control questions, and how they relate to subjective difficulty. The resulting correlation matrix is presented in Table 6, below.
Table 6. Pearson’s $r$ correlation matrix for the ESSAI variables recorded in the ELICIT experiment trials.

<table>
<thead>
<tr>
<th></th>
<th>Tractability</th>
<th>Dynamics</th>
<th>Dependencies</th>
<th>Tractability Control</th>
<th>Dynamics Control</th>
<th>Dependencies Control</th>
<th>Subj. Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tractability</td>
<td>-.02</td>
<td>.03</td>
<td>-.24**</td>
<td>.24**</td>
<td>-.17*</td>
<td>-.33***</td>
<td></td>
</tr>
<tr>
<td>Dynamics</td>
<td>.40***</td>
<td>.42***</td>
<td>-.51***</td>
<td>-.06</td>
<td>.39***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependencies</td>
<td>.23**</td>
<td>-.17*</td>
<td>-.18*</td>
<td>.31***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tractability Control</td>
<td>-.33***</td>
<td>-.14</td>
<td>.45***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamics Control</td>
<td>.22**</td>
<td>-.35***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependencies Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj. Difficulty</td>
<td></td>
<td></td>
<td>-.22**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * correlations are significant at the $p = .05$ level.
** correlations are significant at the $p = .01$ level.
*** correlations are significant at the $p = .001$ level.

Analysis of the correlational relationships between the Endeavor Space dimensions and their respective control questions revealed statistically significant negative correlations between the Tractability variable and its control question, $r(166) = -.24, p = .002$, Dynamics and its control question, $r(166) = -.52, p < .001$, and Dependencies and its control question, $r(166) = -.18, p = .023$. This means that as the response scores on the dimensions increased, the response scores on the dimension control questions decreased, and vice versa. Although the correlations are mostly weak, they point in the intended direction.

Significant correlations were also found between Dynamics and Dependencies, $r(166) = .40, p < .001$, indicating that the two variables move in tandem. However, no significant correlations were found between Tractability and Dynamics, $r(166) = .02, p = .821$, or Tractability and Dependencies, $r(166) = .03, p = .678$.

Furthermore, ratings of subjective difficulty were significantly correlated with Tractability, $r(166) = -.33, p < .001$, meaning that as the tractability of the tasks decreased, the perceived difficulty increased. This result is compatible with current theories on complexity and difficulty, and supports our first hypothesis, H1. Subjective difficulty was also significantly correlated with Dynamics, $r(166) = .39, p < .001$, and Dependencies, $r(166) = .31, p < .001$. This means that as either task dynamics or
dependencies were increased, the tasks were perceived to be more difficult. These results are also in accordance with the literature, and supports our second and third hypotheses, H2 and H3.

### 4.2 ELICIT experiment results

We will now present the performance results and the ESSAI responses from the ELICIT trials. Team performance scores are presented in Table 7, below.

**Table 7. Descriptive statistics of ELICIT team scores.** Possible scores ranged from 0 - 12 points per scenario. N = 8.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Mean (SD)</th>
<th>min - max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>7.00 (2.70)</td>
<td>2.75 - 10.75</td>
</tr>
<tr>
<td>Medium</td>
<td>8.16 (2.37)</td>
<td>4.25 - 10.75</td>
</tr>
<tr>
<td>Hard</td>
<td>6.56 (2.02)</td>
<td>4.00 - 10.00</td>
</tr>
</tbody>
</table>

A Friedman test was conducted to test the differences in team scores between the three scenarios. A non-significant result, $\chi^2(2) = 2.47$, $p = .291$, indicated that team performance scores did not differ significantly between the three scenarios. No post-hoc tests were therefore conducted.

Next, we turn to the results of the ESSAI responses as they relate to the ELICIT scenarios. Table 8 presents the descriptive statistics of the ESSAI data over the Easy, Medium, and Hard ELICIT scenarios.
Table 8. Descriptive statistics of team-level ESSAI responses to the ELICIT scenarios. All scores are between 1 and 7. The ‘Total Complexity’ variable is the aggregate mean of the Tractability, Dynamics, and Dependencies variables. Standard deviations in parentheses. N = 8.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Tractability</th>
<th>Dynamics</th>
<th>Dependencies</th>
<th>Total Complexity</th>
<th>Subj. Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>Easy</td>
<td>5.20 (.06)</td>
<td>3.95 (.78)</td>
<td>4.83 (.72)</td>
<td>4.66 (.53)</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>5.46 (.95)</td>
<td>3.89 (.73)</td>
<td>4.82 (1.01)</td>
<td>4.72 (.59)</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>5.46 (1.00)</td>
<td>4.06 (.76)</td>
<td>4.98 (.76)</td>
<td>4.83 (.55)</td>
</tr>
<tr>
<td>min - max</td>
<td>Easy</td>
<td>1.67 - 7.00</td>
<td>2.17 - 5.88</td>
<td>3.17 - 6.00</td>
<td>3.31 - 5.69</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>2.50 - 7.00</td>
<td>2.25 - 5.50</td>
<td>1.67 - 6.83</td>
<td>2.96 - 5.83</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>2.50 - 7.00</td>
<td>2.38 - 6.38</td>
<td>3.00 - 6.50</td>
<td>3.24 - 5.97</td>
</tr>
</tbody>
</table>

Again, Friedman tests were conducted to test whether ratings of Tractability, Dynamics, and Dependencies differed between the scenarios. The results showed that Tractability ratings did not differ significantly between the Easy, Medium, and Hard scenarios, $\chi^2(2) = 1.00, p = .607$. Ratings of Dynamics also did not differ significantly between the scenarios, $\chi^2(2) = 3.16, p = .206$. Lastly, differences in Dependencies ratings between scenarios also did not reach the level of statistical significance, $\chi^2(2) = 2.39, p = .303$. These results indicate that, at the team level, none of the Endeavor Space dimension factors were rated differently between the scenarios.

Team-level overall complexity between scenarios were also calculated. Scenario total complexity scores—the mean of the three Endeavor Space dimensions—submitted to a Friedman test revealed no significant differences in overall perceived complexity, $\chi^2(2) = 3.25, p = .197$.

Team-level subjective difficulty was also submitted to a Friedman test to calculate differences between scenarios. The results indicated no significant team-level differences in perceived difficulty between the scenarios, $\chi^2(2) = 5.40, p = .064$.

To summarize the above findings; at the team level, no differences were found in any of the three Endeavor Space dimension factors, overall complexity, subjective difficulty, or objective team performance over the Easy, Medium, and Hard scenarios.
Next, we were interested in how Endeavor Space dimension ratings differed between the mid-task and post-task measurement points over the three scenarios. Table 9, below, present descriptive statistics of the dimensions, total complexity, and subjective difficulty as measured mid- and post-task in every scenario.

**Table 9. Mean mid-task and post-task Endeavor Space dimension ratings between scenarios. Standard deviations in parentheses. N = 56.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Checkpoint</th>
<th>Tractability</th>
<th>Dynamics</th>
<th>Dependencies</th>
<th>Total Complexity</th>
<th>Subj. Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>Mid-task</td>
<td>5.03 (.22)</td>
<td>3.88 (.83)</td>
<td>4.92 (.73)</td>
<td>4.61 (.54)</td>
<td>4.70 (1.28)</td>
</tr>
<tr>
<td></td>
<td>Post-task</td>
<td>5.37 (1.10)</td>
<td>4.02 (.91)</td>
<td>4.73 (.95)</td>
<td>4.71 (.62)</td>
<td>4.59 (1.56)</td>
</tr>
<tr>
<td>Medium</td>
<td>Mid-task</td>
<td>5.48 (1.02)</td>
<td>3.83 (.73)</td>
<td>4.79 (1.05)</td>
<td>4.70 (.61)</td>
<td>4.50 (1.43)</td>
</tr>
<tr>
<td></td>
<td>Post-task</td>
<td>5.44 (1.03)</td>
<td>3.94 (.87)</td>
<td>4.86 (1.09)</td>
<td>4.75 (.65)</td>
<td>4.43 (1.48)</td>
</tr>
<tr>
<td>Hard</td>
<td>Mid-task</td>
<td>5.52 (1.01)</td>
<td>4.04 (.78)</td>
<td>4.90 (.83)</td>
<td>4.82 (.58)</td>
<td>4.79 (1.11)</td>
</tr>
<tr>
<td></td>
<td>Post-task</td>
<td>5.41 (1.14)</td>
<td>4.08 (.87)</td>
<td>5.05 (.79)</td>
<td>4.85 (.58)</td>
<td>5.05 (1.18)</td>
</tr>
<tr>
<td>Overall</td>
<td>Mid-task</td>
<td>5.34 (.84)</td>
<td>3.92 (.66)</td>
<td>4.87 (.74)</td>
<td>4.71 (.49)</td>
<td>4.66 (1.02)</td>
</tr>
<tr>
<td></td>
<td>Post-task</td>
<td>5.41 (.95)</td>
<td>4.01 (.76)</td>
<td>4.88 (.74)</td>
<td>4.77 (.54)</td>
<td>4.69 (1.06)</td>
</tr>
</tbody>
</table>

Two-way repeated measures analyses of variance (ANOVA) were conducted for ratings of each Endeavor Space dimension, total complexity, and subjective difficulty. The independent variables were scenario (three levels: Easy, Medium, Hard) and checkpoint (two levels: mid and post). Tukey corrections were used for all post hoc tests. For Tractability, the interaction effect of scenario and checkpoint was significant, $F(2, 110) = 5.19, p = .007, \eta^2_p = .086$. Post hoc tests revealed significant differences between the mid- and post-task measurement points for the Easy scenario, $t(159) = -2.97, p = .039$, suggesting that participants reported the Easy scenario to be more tractable once they had completed it compared to when they were still playing it. Also, the mid-task Tractability ratings differed between the Easy and Hard scenarios, $t(184) = -3.39, p = .011$, meaning that participants rated the Hard scenario to be more tractable compared to the Easy scenario, as reported mid-task. No other significant differences were found regarding the Tractability dimension. For the Dynamics dimension, the interaction effect of scenario and checkpoint was not significant, $F(2,$
dimension, a Greenhouse-Geisser correction was used due to the assumption of sphericity being violated. Here, the interaction effect of scenario and checkpoint was significant, $F(1.77, 97.15) = 3.45, p = .041, \eta^2_p = .059$, but post hoc test revealed no pairwise differences. Furthermore, no significant interactions between scenario and checkpoint were found for total complexity, $F(2, 110) = .44, p = .648, \eta^2_p = .008$, or subjective difficulty, $F(2, 110) = 2.07, p = .131, \eta^2_p = .036$, meaning that mid- and post-task ratings of neither of these variables differed within or between scenarios.

Next, Friedman tests were conducted to investigate whether responses to individual ESSAI items differed between the Easy, Medium, and Hard scenarios. Table 10, below, provides descriptive statistics of the items over each scenario.

**Table 10. ESSAI item responses over the three ELICIT scenarios. Scores represent means of mid-task and post-task ratings. Standard deviations in parentheses. N = 56.**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (SD)</td>
<td>min - max</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tractability</td>
<td>Q1</td>
<td>5.54 (1.23)</td>
<td>5.73 (1.10)</td>
<td>5.88 (1.03)</td>
<td>1.5 - 7</td>
<td>1.5 - 7</td>
<td>2.5 - 7</td>
</tr>
<tr>
<td></td>
<td>Q2</td>
<td>5.52 (1.18)</td>
<td>5.78 (1.07)</td>
<td>5.71 (1.07)</td>
<td>2 - 7</td>
<td>2 - 7</td>
<td>2 - 7</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>4.55 (1.22)</td>
<td>4.87 (1.13)</td>
<td>4.79 (1.37)</td>
<td>1.5 - 7</td>
<td>3 - 7</td>
<td>1 - 7</td>
</tr>
<tr>
<td>Dynamics</td>
<td>Q5</td>
<td>3.51 (1.29)</td>
<td>3.43 (1.36)</td>
<td>3.70 (1.36)</td>
<td>1.5 - 7</td>
<td>1 - 6</td>
<td>1 - 6.5</td>
</tr>
<tr>
<td></td>
<td>Q6</td>
<td>2.98 (1.17)</td>
<td>2.72 (1.06)</td>
<td>3.01 (1.20)</td>
<td>1 - 6.5</td>
<td>1 - 5.5</td>
<td>1.5 - 7</td>
</tr>
<tr>
<td></td>
<td>Q7</td>
<td>4.90 (1.01)</td>
<td>5.01 (1.20)</td>
<td>4.94 (1.15)</td>
<td>2.5 - 7</td>
<td>2.5 - 7</td>
<td>2.5 - 7</td>
</tr>
<tr>
<td></td>
<td>Q8</td>
<td>4.40 (1.36)</td>
<td>4.39 (1.42)</td>
<td>4.60 (1.40)</td>
<td>1.5 - 7</td>
<td>1 - 7</td>
<td>1 - 7</td>
</tr>
<tr>
<td>Dependencies</td>
<td>Q10</td>
<td>5.38 (1.33)</td>
<td>5.17 (1.55)</td>
<td>5.52 (1.57)</td>
<td>1.5 - 7</td>
<td>2 - 7</td>
<td>1 - 7</td>
</tr>
<tr>
<td></td>
<td>Q11</td>
<td>3.21 (1.24)</td>
<td>3.38 (1.37)</td>
<td>3.18 (1.41)</td>
<td>1.5 - 6</td>
<td>1 - 6.5</td>
<td>1 - 6</td>
</tr>
<tr>
<td></td>
<td>Q12</td>
<td>5.88 (1.16)</td>
<td>5.92 (1.20)</td>
<td>6.23 (0.83)</td>
<td>2 - 7</td>
<td>1 - 7</td>
<td>3 - 7</td>
</tr>
</tbody>
</table>

There was a statistically significant difference in responses to Question 12 between scenarios, $\chi^2(2) = 6.00, p = .049$. A post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni-corrected significance level set at $p < .017$. This
revealed no significant differences in responses to Question 12 between the Easy and Medium scenarios, $Z = -1.17, p = .865$, or the Easy and Hard scenarios, $Z = -2.25, p = .024$. However, a significant difference in mean Q12 responses was found between the Medium and Hard scenarios, $Z = -2.55, p = .011$. This means that participants reported feeling more affected by each other's tasks and actions in the Hard scenario compared to the Medium scenario. No other significant differences were found.

Lastly, Pearson’s $r$ correlations were analyzed to investigate the relationships between ELICIT performance and Tractability, Dynamics, Dependencies, overall complexity, and subjective difficulty. For this analysis, the ELICIT scores were normalized over the roles to represent relative scores to compensate for the fact that the top manager could receive four points whereas the other roles could only score two or one point(s). This procedure involved averaging the scores on the subproblems (i.e. who, what, where, when) assigned to each role so that scores would range between 0 and 1 for all roles, where 0 represents the worst possible score relative to the role and 1 representing the best possible score relative to the role. The resulting correlation matrix is presented in Table 11, below.

Table 11. Pearson’s $r$ correlation matrix of ESSAI variables recorded in the ELICIT scenarios, along with role relative ELICIT performance scores.

<table>
<thead>
<tr>
<th></th>
<th>Tractability</th>
<th>Dynamics</th>
<th>Dependencies</th>
<th>Total Complexity</th>
<th>Subj. Difficulty</th>
<th>ELICIT Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tractability</td>
<td>-.02</td>
<td>.03</td>
<td>.66***</td>
<td>-.33***</td>
<td>-.31***</td>
<td>.32***</td>
</tr>
<tr>
<td>Dynamics</td>
<td>.40***</td>
<td>.64***</td>
<td>.39***</td>
<td>-.22**</td>
<td>.13</td>
<td>.04</td>
</tr>
<tr>
<td>Dependencies</td>
<td>.70***</td>
<td>.31***</td>
<td>-.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj. Difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.31***</td>
<td></td>
</tr>
<tr>
<td>ELICIT Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * correlations are significant at the $p = .05$ level. ** correlations are significant at the $p = .01$ level. *** correlations are significant at the $p = .001$ level.
All three Endeavor Space dimensions correlated significantly with total complexity. This is entirely unsurprising since total complexity is a construct of the three dimensions. As such, these results are meaningless and will not be discussed further.

Performance outcome in ELICIT was significantly correlated with Tractability, $r(166) = .32, p < .001$, indicating that an increase in perceived tractability was associated with better performance. This supports our fourth hypothesis, H4. ELICIT performance was also significantly correlated with Dynamics, $r(166) = -.22, p = .004$. This suggests that as tasks are perceived to be more static and stable, performance tend to increase. This result lends support to our fifth hypothesis, H5. However, no significant relationship was found between ELICIT performance and Dependencies, $r(166) = -.11, p = .146$, a result which cannot reject our sixth null hypothesis, H6. Finally, ELICIT performance was significantly correlated with subjective difficulty, $r(166) = -.31, p < .001$, indicating that as tasks were perceived to be more difficult, performance outcomes would decrease.
5 Discussion

In this chapter, we will first discuss our methodological choices and design decisions before we go on to discuss the implications of our findings.

5.1 Method

Regarding the Pandemic trails, it is worth noting that the changes made in the Pandemic scenarios to manipulate the Endeavor Space dimensions were highly hypothetical. As there exists no theoretical background on Endeavor Space implementation of the dimension in boardgames, the manipulations that were made had to be theoretically inferred, e.g. that increasing the number of epidemic cards (which introduces new challenges to the game) should increasingly evoke the feeling that the game environment is dynamic. Also, Pandemic may be more open to manipulation on some dimensions than others. For instance, it is easier to manipulate dynamics as Pandemic is, by its very design, a dynamic environment than it is to manipulate tractability. It is possible that more precise manipulations could have been made had more time been given to the development of these manipulations. However, given the limited time available for pretesting the ESSAI tool—and lack of pre-existing knowledge of how Endeavor Space dimensions manifest in boardgames—it was deemed unnecessary to do any in-depth analysis of how Pandemic components relate to the dimensions. The primary focus of the Pandemic trails was to test the ESSAI tool in an interactive environment which could manifest the dimensions, and thus the manipulations were considered good enough.

As mentioned in the ESSAI design rationale section (see section 3.1.1), the choice was made to formulate the questionnaire items to be applicable to multiple domains. Thus, the focus was on how the items could probe the dimensions in relation to generic task characteristics rather than being specific to how the dimensions may manifest in ELICIT. The ESSAI thus follows the rationale used for formulating the Endeavor Space dimensions and aligns with the intended purpose of the Endeavor Space to be applicable to any kind of endeavor. However, this rationale has its drawbacks; if the formulations are generic, they are also more open to interpretation by the participant. It is therefore possible that individual, personal characteristics may influence a respondent’s ability to answer the questionnaire as intended. For example, a
participant with knowledge of system theory may, due to their background, be more adept at reasoning around system components and how they interact, enhancing their ability to infer and reason about such relations when prompted in a questionnaire. Conversely, if a participant has no relevant prior knowledge, the questions may be misinterpreted. The influence of participants’ personal factors on results is, however, always a risk factor associated with subjective measuring tools, and is not unique to the ESSAI. The current study has tried to guard against this as best as possible by aiming to formulate the items in ESSAI as clearly as possible, refraining from using context dependent words, e.g. “system”, which requires some knowledge of what a system is (or could be) to be relatable. Questionnaires are best developed through an iterative process where multiple formulations are tested until one set of questions is decided upon. Unfortunately, due to the time constraints associated with this study, it has not been possible to properly test several iterations of the ESSAI tool to see which formulations are the most communicative. However, considering that two pre-tests (inhouse at FOI and the Pandemic trails) using the ESSAI were conducted, it is the opinion of the authors that the items in the latest iteration of the ESSAI are reasonably understandable and concise. As a final thought on the ESSAI items themselves, it is possible that their general nature did not fit how some aspects of the Endeavor Space were instantiated in ELICIT. For instance, the issues with Question 11—which asks respondents to rate how sequential they perceive their tasks to be—could be due to the fact that there is no way to directly manipulate task sequentiality in ELICIT. The ability of the Top Manager to solve his or her task(s) is entirely decoupled from that of the workers. It does not matter to the Top Manager whether or not the When-worker solves their problem. What matters to the Top Manager (or any team member, for that matter) is the ability of his or her subordinates to share and funnel information through the hierarchy pipeline. In ELICIT, (human) members are not dependent on how well their colleagues solve their “riddles”, but how well they communicate. This (perceived) need to communicate is probed by Question 10. Therefore, Question 11 serves no purpose in ELICIT, but for the ESSAI to be a generalized tool to capture Endeavor Space assessments, Question 11 is still needed. Also, the ESSAI questionnaire would likely benefit from a question directly gauging the respondents perceived temporal pressure as a part of assessing the dynamics dimension.
Another note on the ESSAI; the decision to use Likert-type scales to rate the items was not an entirely successful one. Although Likert-type scales are commonly used in self-assessment instruments, they do not completely lend themselves to quantitative statistical analyses. For instance, Likert data tend to exhibit skewness or kurtosis. To allow parametric analyses, these data would have to be transformed to satisfy the numerous assumptions associated with such tests. This makes interpretation of the results more difficult. Another option is to use non-parametric techniques, as was the case in the current study. However, these tests are typically not as flexible or as powerful as their parametric counterparts. Furthermore, Likert data are coarse-grained in the sense that they only deal with natural numbers. For our purposes, real numbers would have been more appropriate and useful. In averaging ordinal (natural numbers) data, we did get real numbers, but these were still constrained by the ordinal nature of their Likert-type antecedents. It is our opinion and belief that the ordinal Likert-type scale used to rate the ESSAI items should be replaced with a thoroughly continuous rating scale. The new scale could be something as trivial as a straight line with semantic anchors at each end, such as the current *Entirely Disagree* and *Fully Agree* anchors. Respondents would be asked to mark their answer on the line, and the researcher would use a ruler to measure the distance in millimeters from the starting point to the (center of) the response marking. The distance would represent the “score”. This scale would be much more fine-grained than the current scale. It is likely that analyses of data recorded with this new scale would be easier to perform and reveal finer details in the underlying phenomena being studied.

Next, it was decided in the experiment design that the teams participating in the study would not receive any training. Participants would be introduced to ELICIT at the beginning of each experiment session and be allowed to ask questions, but no separate test runs or extensive training sessions were held. This may have impacted a couple of factors which may have influenced the results in the study. First, training the participants in ELICIT may have resulted in better ELICIT performances. In the past, ELICIT has predominantly been used for conducting large-scale agent-based simulations and has therefore seen little change in its user interface over the years. Participants in the study often commented on how unhelpful the ELICIT interface was, e.g. not being able to eliminate duplicates, or the process of sending and sorting
information being cumbersome in the current interface. Braarud and Karwin (2011) marks *interface complexity*—how well an operator's instruments support them in their task—as a contributing factor to overall task complexity. It is likely that familiarizing the participants with the ELICIT interface over the course of a training session or scenario could have improved their ability to formulate strategies based of the capabilities afforded by the ELICIT interface. As this was not the case in the current experiment design, the scenario which participants encountered first essentially became their training scenario and was therefore likely experienced as difficult regardless of the hypothesized scenario difficulty level.

Another aspect which could be argued to have been influenced by the lack of training is the teams’ overall cohesion. It was not set as a prerequisite for this study that the participating teams would have to be professional in any capacity. Thus, it was understood that the impromptu nature of the teams would affect their ability to cooperate. It is well-documented that the effectiveness and performance of a team is largely reliant on its cohesion, which can be developed and enhanced through training (Cooke et al., 2007, 2004). Although the ELICIT performance of each team was of secondary concern for the current study, a team’s ability to cooperate does impact each participant’s capacity to complete their personal tasks, likely impacting their perception of difficulty and the tasks themselves. Given that a team performs poorly, the tasks may be perceived as more complex and difficult than they actually are. It is therefore possible that trained teams would have exhibited different behaviors in ELICIT and therefore ranked the ESSAI dimensions differently. Again, this is a general drawback associated with subjective measurements of any kind - there will always be uncontrollable factors that influence subjective perception. However, it would be interesting to conduct similar experiments with trained teams to investigate potential differences in performance and general behavior.

As a final thought on ELICIT as a research platform, it is the belief of the authors that tasks in ELICIT are mainly ones of processing streams of information, and do not, generally speaking, give rise to strong emotions or experiences. As a domain, ELICIT is not relatable to ordinary people - it can appear as “artificial” in the sense that it carries little emotional meaning to participants. The tasks themselves require or evoke no emotional engagement. This means that experiential reactions to events in ELICIT
are not strong to begin with, meaning that differences in experiences between our three scenarios were relatively tiny. Consider the following as it relates to surprising events and the Dynamics dimension; the Who-worker suspects that the magenta group will be the perpetrator, but receives a factoid saying that all members of the magenta group are in custody. This revelation would be more likely to evoke a mundane reaction, e.g. “oh, I see”, rather than a strong and emotional reaction in terms of assessing the magnitude of this change in circumstance. For comparison, consider a more relatable simulated task where the participant was given the role and responsibility of making sure that a burning building was entirely evacuated. Imagine their reaction upon learning that there was still a child inside. Such emotional investment and engagement are better evoked in live simulations than in virtual simulations like ELICIT. It is possible that this lack of immersion results in mild experiences overall, whereby any potential differences between experiences become small and difficult to measure. This could possibly have been remedied by providing a more relatable and emotionally meaningful context to our participants. By incorporating elements of live simulation, e.g. narrative contexts, meaningful roles, or impactful consequences, the ELICIT environment could have been made more engrossing and immersive. In such a setup, ELICIT would be more akin to a hybrid simulation than a strictly virtual one. In this way, the experience of interacting and working in ELICIT would have been enhanced, and differences between experiences could have been more easily detected. Also, another factor adding to the lack of immersion may stem from the experiment design, in particular the background of the recruited participants. The current study used teams consisting of university students with varying backgrounds, and their ability to relate to the tasks in ELICIT was therefore limited. Constructing teams with trained professionals from relevant fields, e.g. military personnel, intelligence workers etc., may see an increased sense of immersion due to their familiarity with the types of tasks available in ELICIT. It is also possible that participant immersion would benefit from adding competitive aspects to the experiment design. Participants could be encouraged to immerse themselves thoroughly in their task by having teams compete against each other in some way, e.g. having the best performance score. In the current study, most teams were observed to display competitive behaviors, often being keen on finding out their results and discussing strategies for improving their performance and so on.
5.2 Results

We will begin this results discussion by reviewing the ESSAI tool itself. Factor analyses conducted to assess the internal consistency of the ESSAI produced mixed results. First, the analyses provide tentative support to the idea that the ESSAI items measure three distinct constructs rather than probing the same underlying one. Specifically, the items representing the Tractability and Dynamics dimensions fit well together in their respective Endeavor Space dimension factors, with the exception of Question 7 showing a stronger loading on the Tractability factor than its intended Dynamics factor. Question 7, after all, essentially asks whether changes were tractable. This item, therefore, needs to be reworked. On the other hand, the Dependencies dimension is less cohesive, with Question 11 being a particularly bad fit with the other two Dependencies items, and Question 10 being affected by the other dimensions. Therefore, Question 11, which probes the feeling of tasks being sequential, and Question 10, which probes the need for communication, miss their marks in some way. Further iteration is needed to bring Question 10 and Question 11 in line with Question 12 in the Dependencies factor. Furthermore, reliability analyses showed that the Tractability factor was reasonably consistent and reliable. Its Cronbach’s α of .81 was well beyond the typical thresholds of .7 or .8, suggesting that the Tractability items were adequately communicative and understandable, needing only minor adjustments. The items constituting the Dynamics factor were less successfully designed. With an α value of .64, the Dynamics factor was below even the .7 threshold. Thus, the Dynamics dimension items need further iteration. Lastly, the reliability analyses results echoed those of the validity tests regarding the Dependencies dimension items, its α value of .28 indicating major reliability issues. Removing the ill-fitting Question 11 would still only increase the reliability of the Dependencies factor to .54, which is well below satisfactory levels. The Dependencies dimension items, therefore, need more work and iteration. For instance, Question 10 is likely very domain-dependent. Communication in ELICIT does not refer to typical forms of communication, such as talking or typing, but manifest solely as sharing and receiving prepared packets of information. It is possible that this activity was interpreted as a task rather than “communication” in the colloquial sense. As previously discussed, the bad fit and reliability of Question 11 could be due to its vagueness in relation to
problems in ELICIT. The intended interpretation is that communication—i.e. receiving and distributing factoids—is included in the “tasks” clause, the idea being that tasks needed to be completed—i.e. factoids being shared—to enable the completion of other tasks, such as identifying and submitting problem solutions. It is likely that this was not the interpretation of our participants.

We will now discuss the results as they relate to our original research questions.

**RQ1. How can the Endeavor Space dimensions be instantiated and manipulated in the ELICIT simulator?**

The Endeavor Space, as interpreted in the current study, consists of three dimensions; Tractability, Dynamics, and Dependencies. To answer research question RQ1, three scenarios were created in which the dimensions had been manipulated to elicit three different experiences of the Endeavor Space dimensions. We will account for each dimension and its respective manipulations below and discuss how successfully they were implemented in ELICIT.

**Tractability**

Tractability was interpreted as the degree to which one can gain an overview and understanding of the constituent parts of a problem and how they interact. Low tractability would thus indicate that it is difficult to gain an overview and understanding of a problem, while high tractability would imply the opposite. The tractability of a problem in ELICIT would thus be a consequence of manipulating how sizeable the problem is, i.e. the number of parts that constitute the problem, and the degree to which it was possible to discern a component as relevant to the problem. Tractability was therefore manipulated through the number of factoids present in a given scenario (the size of the problem) and the signal-to-noise ratio amongst those factoids (the ratio of relevant to irrelevant problem components). The theorized effects were that as tractability was lowered by reducing the signal-to-noise ratio and increasing the number of factoids, participants would perceive the task as more difficult and complex as it became more difficult to grasp and assess the problem.

Overall, it is the opinion of the authors that the tractability manipulations (in theory) were relevant to the theoretical construct of tractability. Varying the
number of components to keep track of, and manipulating the opaqueness of component connection by introducing noise, both should cause a problem to seem more complex and more difficult. However, it is uncertain how well these manipulations translate into subjective experience of tractability.

The manipulations available in ELICIT are primarily quantitative in nature, such as increasing the number of components or increasing number of noise elements. It is not obvious that quantitative manipulations of this kind translate well into subjective experiences. In other words, the feeling of complexity—as stemming from the subjective understanding of a problem—may be entirely unrelated to how objectively large or opaque a problem is. This relates to the subjective aspect of complexity (Braarud & Kirwan, 2011; Haerem et al., 2015; McIntyre, 1998) as explained in section 2.2.1. Expertise (or a lack thereof) may affect how participants feel they understand the problem regardless of its “objective” size. A participant who is experienced and knowledgeable may find the problem uncomplex even though the problem is objectively rated as complicated and complex. A novice participant may also perceive a complex problem as uncomplex but rather due to misunderstanding the problem or not being able to fully understand the situation, thus misjudging its complexity.

**Dynamics**

Dynamics was manipulated through the time pressure inflicted on the participants by the scenario time (15 minutes), the total number of factoid waves that was delivered, and the frequency with which the factoid waves were delivered. However, as mentioned in section 3.2.6, these three manipulations were kept constant across all three ELICIT scenarios. This will be explained further here.

The decision to keep the frequency and number of factoid waves at a constant was because, in the ELICIT version used for the current study, it was not possible to create more factoid sets nor change their distribution in any meaningful way. Dynamics was therefore a somewhat difficult dimension to manipulate as the settings available in ELICIT does not allow much variability regarding how factoid waves are distributed. Factoid waves can be varied in their frequency—their number and how often waves are distributed—but they will always follow a set pattern, i.e. the time interval may be
adjusted to deliver a wave each minute, or a wave every other minute and so on, but the waves may not be programmed individually. Thus, it is not possible to set one wave at one minute, one wave at the fourth minute, and another wave at the fifth minute, and so on. This limits the possibility to create dynamic events as waves are only distributed at set intervals. Allowing for individual programming of the delivery time for each factoid distribution wave could improve the capacity for manipulating dynamics in ELICIT scenarios.

Scenario time was kept at a constant due to its impact on difficulty being unclear. However, it was believed that the time pressure aspect would differ indirectly across the scenarios due to interactions between the Tractability and Dependencies dimensions. It was argued that allowing for more time in the Hard scenario would negatively impact its difficulty level and dynamics. The reasoning for this was that if one were to perform a difficult task but had adequate time to do so, the task may not be perceived as difficult since it could be completed at a leisurely pace. Time pressure was thus seen as the product of indirect manipulations stemming from the interaction between available time (dynamics), tractability (size of problem), and problem-solving capabilities (dependencies). The Hard scenario would require more work to be done per available scenario minute in order to complete the tasks compared to the Easy and Medium scenarios, thus increasing applied time pressure.

Overall, it is the belief of the authors that the dynamics dimension is not well captured by the ELICIT manipulations, and that the manipulations do not effectively translate into subjective experiences of dynamics. Although frequency of factoid wave distribution and number of factoid waves—in theory—translates to number and frequency of events in an endeavor, it is unclear that such manipulations are sufficient to create effects dramatic enough to elicit different subjective experiences for the participants. Much like the manipulations for the Tractability dimension, this is most likely due to the inability of quantitative manipulations to evoke qualitative experiences. While likely that all participants perceived that changes were occurring in ELICIT, it is doubtful that these changes evoked any subjective experiences of dynamics since the changes were lacking in meaningful content, surprise factor, and magnitude (a sense that the change was significant).
Dependencies

Dependencies was manipulated in ELICIT through varying the amount of non-organic information available to each participant. The amount of non-organic information available would directly influence the degree to which the teams would have to rely on communication and cross-level information sharing. Decreasing the amount of non-organic information needed would decrease the need to distribute information as each participant would have the information needed for them to complete their task. Increasing the amount of non-organic information needed decreases the need for cross-level information sharing, as each participant directly receives the information that is needed to complete their task. Thus, manipulating the amount of non-organic information available is—in theory—an aspect of manipulating how dependent each participant is on the organization it is working in. In the Hard scenario where dependencies were intended to be high, it was thus argued that participants were dependent on each other to solve their respective tasks as information would have to be shared extensively for each participant to receive the information they needed to solve their task.

Overall it is the belief of the authors that this manipulation does correspond well with the theoretical construct of dependencies. However, it should be noted that dependency between tasks is only indirectly manipulated through this manipulation. ELICIT allows for no direct manipulation of sequentiality of tasks. Therefore, although the participants are dependent on one another to share information, there is no direct dependency between tasks. As previously discussed in section 5.1, the top manager can complete his or her task entirely without regard for how others in the organization has completed their tasks. That is, a direct manipulation of task dependency would be to create, in ELICIT, a scenario where the top manager could only complete his or her tasks given that the workers and middle managers working on similar problems had completed their tasks. This notion is further supported by the fact that Q11 ("I felt that tasks needed to be completed in sequence, one after another") in the ESSAI tool was marked as non-contributing to the dimension of dependency. It is therefore safe to conclude that task dependency is not properly manipulated in the current iteration of the ELICIT scenarios.
RQ2. How do manipulations of Endeavor Space dimensions manifest as subjective experiences?

To answer this research question, we analyzed differences in mean ESSAI response scores between the Easy, Medium, and Hard scenarios. At the team level, no such differences were found. This is unsurprising considering this study’s limited sample size of only eight teams. It is possible, perhaps even likely, that the differences in subjective effects caused by our Endeavor Space dimension manipulations are too small to be statistically detected with so few teams. As previously discussed, ELICIT—as a domain—is somewhat undramatic. It could be that ELICIT is unable to evoke strong feelings and experiences and is rather perceived as a generally monotonous activity. If that is the case, it would require our ELICIT scenarios to be wildly different from each other for any potential differences in subjective experiences to emerge. Given our results, these differences—if they exist—are too small and nuanced to be detected with the current sample size. However, the overall pattern follows the expected trajectory; Dynamics, Dependencies, overall complexity, and subjective difficulty (with the notable exception of Tractability) all indicate increased response scores in the Hard scenario compared to the Easy scenario. It is the opinion of the authors that, given more data, this general pattern would expand and grow in salience. Furthermore, it is possible that the current sample size could have been enough if a different problem domain had been used, one which affords more visceral experiences that are more easily discernible and measured.

RQ3. How do manipulations of Endeavor Space dimensions affect how individuals perceive difficulty, and how does it impact their performance?

We found that Tractability, Dynamics, and Dependencies all correlate with perceived difficulty as predicted in the available literature (see e.g. Braarud and Kirwan, 2011): low Tractability was associated with an increased sense of difficulty, as were high Dynamics and high Dependencies. Furthermore, high ratings of perceived difficulty (and low Tractability and high Dynamics) were associated with reduced ELICIT performance. Although the Easy, Medium, and Hard scenarios did not differ significantly in terms of perceived difficulty or ELICIT performance, the overall pattern suggests that the Easy scenario—which was designed to be high in Tractability and low in Dynamics and Dependencies—
saw enhanced ELICIT performance and reduced levels of perceived difficulty in comparison to the Hard scenario. This pattern, when considered alongside the significant relationships between perceived difficulty and the three Endeavor Space dimensions, further support the previously discussed notion that more data would have confirmed the effects of our manipulations. As it stands, however, we can only conclude that Endeavor Space ratings did correlate with perceived difficulty and objective performance alike, but we cannot confirm that our three scenarios differed in this regard.
6 Conclusions

The current study set out to investigate the potential for using the ELICIT simulator as a tool for implementing the Endeavor Space dimensions (Tractability, Dynamics, and Dependencies). To accomplish this, the current study instantiated the dimensions by theoretically mapping each dimension to corresponding configurable settings in ELICIT. Based of these theoretical mappings, three ELICIT scenarios were developed. These were argued to correspond to different regions of the Endeavor Space due to differences across the three dimensions: one Easy scenario (high tractability, low dependencies, low dynamics), one Medium scenario (medium tractability, medium dependencies, medium dynamics), and one Hard scenario (low tractability, high dependencies, and high dynamics). To verify that the three ELICIT scenarios differed across the dimensions as intended, the Endeavor Space Subjective Assessment Instrument (ESSAI) tool was developed to capture perceived complexity as constituted by the three dimensions. An experimental study was then conducted in which eight teams of seven participants played each of the three ELICIT scenarios and rated their complexity using the ESSAI. Two sets of statistical analyses had to be conducted based on the gathered data: one to investigate validity and reliability of the ESSAI tool, and the second to analyses how teams had performed in and assessed the scenarios.

The primary results—those regarding the success of instantiating the Endeavor Space dimensions in ELICIT—were inconclusive. No significant results were found regarding differences between how the three scenarios had been perceived across the Endeavor Space dimensions (as measured by the ESSAI). This indicates that the manipulations performed in ELICIT, i.e. the three scenarios, failed to produce significantly different experiences of complexity. Thus, no conclusions can be drawn in regard to either of the three research questions. However, the results supported the rejection of the null hypothesis for five of our six hypotheses - our sixth null hypothesis could not be rejected. In other words, all subjective assessments of the dimensions were connected to the subjective experience of difficulty and objective ELICIT performance, except that perceived Dependency was unrelated to ELICIT performance. This supports most of the predictions made in the literature and suggests that the findings of the current study are still incomplete.
6.1 Future research

The conclusions of this study hints to several improvements to the ESSAI tool itself and to how research on Endeavor Space ought to be conducted in the future. We will now discuss each of these in turn.

The ESSAI tool was designed to be a generalized instrument to assess endeavor complexity. However, it has still only been applied to assess two systems: Pandemic and ELICIT. To further establish the validity and reliability of ESSAI, it would need to be applied in more, and substantially different, domain areas. An example would be to replicate the current study in the microworld C3Fire, which in theory also could be used to create scenarios according to the Endeavor Space dimensions but be instantiated in a task (firefighting) significantly different from ELICIT. This would serve to polish and refine the ESSAI items to ensure their general relevance and interpretability. Furthermore, a qualitative study could be conducted where participants engage and interact with a system and are later asked in interviews about how they would describe their experiences as they relate to the Endeavor Space dimensions. These insights could further point to strengths or issues with the ESSAI items.

Future studies should also investigate whether and how trained teams have different perceptions of complexity (or tractability, dynamics, or dependencies) than do untrained teams. Team cohesion and experience have mitigating effects on perceived workload and difficulty. It therefore seems likely that trained teams would report lower assessments of complexity while, at the same time, exhibit higher levels of efficiency and performance outcomes.

Furthermore, the current study did not focus on the complexity ratings of individual roles. This would be an important consideration in future research. Different roles and responsibilities come with different workloads and challenges. Understanding how these factors interact is an important step in exploring and specifying the Endeavor Space dimensions.

Our final remark ties into the above point and our discussions about how the Endeavor Space manipulations can be manipulated directly or indirectly. The dimensions are not entirely discrete and separate from each other - they overlap and interact in different
ways. To fully understand the intricacies of the Endeavor Space, these interactions and relationships need to be charted.
7 Bibliography


American Psychological Association. https://doi.org/10.1037/10690-005


Appendix A – Endeavor Space Subjective Assessment Instrument (ESSAI)

This questionnaire is designed to capture your subjective assessment of the problem or task you and your team has been assigned.

When filling out the form, please do not focus too much on singular events. Instead, try to consider your overall impression of the problem. Be aware that your performance will influence your assessment. Try to assess the problem itself regardless of how well you are currently doing.

Please read each item carefully. Answer each item by drawing a circle around the number that best represents your level of agreement with each statement.

1. I understood how the problem was structured, i.e. its component elements or how they interacted.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

2. I understood what actions to take given my understanding of the problem.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

3. Given my understanding of the problem, I could anticipate how it would change over time.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

4. I felt that the problem -and our task as a team- was complex and difficult to understand.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

5. I often had to re-evaluate and adapt my planned course of action.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
6. When circumstances changed, I felt that they changed drastically, so I had to entirely change my plans rather than adjust them.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

7. When I had to adjust my plans and actions, I felt that I could do so easily.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

8. I felt that circumstances were prone to change.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

9. I felt that the circumstances of my tasks remained static, so I didn’t need to adapt my plans and actions.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

10. I felt that I needed to communicate extensively with others to complete my tasks.

    | Entirely Disagree | Fully Agree |
    |-------------------|------------|
    | 1                 | 2          |
    | 3                 | 4          |
    | 5                 | 6          |
    | 7                 |            |

11. I felt that tasks needed to be completed in sequence, one after another.

    | Entirely Disagree | Fully Agree |
    |-------------------|------------|
    | 1                 | 2          |
    | 3                 | 4          |
    | 5                 | 6          |
    | 7                 |            |

12. I felt that I was affected by the tasks and actions performed by others in my organization.

    | Entirely Disagree | Fully Agree |
    |-------------------|------------|
    | 1                 | 2          |
    | 3                 | 4          |
    | 5                 | 6          |
    | 7                 |            |
13. I felt that tasks were independent from each other.

<table>
<thead>
<tr>
<th>Entirely Disagree</th>
<th>Fully Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

14. Overall, I think that the scenario or task -in its entirety- was...

<table>
<thead>
<tr>
<th>Very Easy</th>
<th>Very Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Please double-check that you have provided answers to each item.
I samverkan med FOI utföra vi denna studie med syfte att utreda komplexa system.

Spelsessionen kommer ta ca 1-1.5 timme där ni kommer få spela två ronder i brädspelet Pandemic. Varje rond tar max 30 minuter, viss tid kan tillkomma på grund av förklaring av regler och utfyllnad av enkäter. Du kommer få spela tillsammans med 3 andra spelare. Tanken är att ni spelar spelet fritt under 30 minuter tills ni antingen vinner, förlorar, eller tiden tar slut. När ni har avslutat varje spelrond kommer du att få fylla i ett antal enkäter med frågor om hur ni upplevde spelandet. Det är därför inte er prestation i spelet som undersöks men vi rekommenderar att ni spelar för att vinna.

Det lagrade datamaterialet kommer att vara helt anonymt, och lagras så att inga utomstående har åtkomst till det. Det insamlade datamaterialet kan komma att användas i andra forskningsprojekt i framtiden (fortfarande helt anonymiserat), men aldrig för några kommersiella syften. Resultaten kommer att presenteras på ett sådant sätt att ingen utomstående kommer att kunna koppla informationen direkt till dig – inga namn kommer att nämnas.

Vi har tyvärr ingen möjlighet att kompensera dig för din medverkan i studien.

Din medverkan i studien är givetvis helt frivillig, och du har både rätt och möjlighet att när som helst och utan risk för negativa påföljder för dig–avbryta ditt deltagande och begära att dina data raderas eller förstörs.

Jag (deltagaren) intygar att jag har läst och förstått ovanstående information och ger härmed mitt informerade samtycke att delta i studien.

______________________________  ________________
Underskrift  Datum (åååå-mm-dd)
Appendix C – ELICIT trials consent form

Projektledare: Jacob Weilandt, mastersstudent i kognitionsvetenskap vid Linköpings universitet, Oscar Bjurling, mastersstudent i kognitionsvetenskap vid Linköpings universitet

Handledare: Björn Johansson, Totalförsvarets forskningsinstitut (FOI)

I samverkan med FOI utför vi denna studie med syftet att studera komplexa system.


Det lagrade datamaterialet kommer att vara helt anonymt och lagras så att inga utomstående har åtkomst till det. Det insamlade datamaterialet kan komma att användas i andra forskningsprojekt i framtiden (fortfarande helt anonymiserat) men aldrig för några kommersiella syften. Resultaten kommer att presenteras på ett sådant sätt att ingen utomstående kommer att kunna koppla informationen direkt till dig – inga namn kommer att nämnas.

Din medverkan i studien är givetvis helt frivillig, och du har både rätt och möjlighet att när som helst–och utan risk för negativa påföljder för dig–avbryta ditt deltagande och begära att dina data raderas eller förstörs.

Som tack för ditt deltagande får du en biobiljett!

Jag (deltagaren) intygar att jag har läst och förstått ovanstående information och ger härmed mitt informerade samtycke att delta i studien.

Underskrift

Datum (åååå-mm-dd)

101
Appendix D – NASA Task Load Index (TLX)

Mark your answer to each question by filling in the empty circle.

<table>
<thead>
<tr>
<th>Mental Demand</th>
<th>How mentally demanding was the task?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Low</td>
<td></td>
</tr>
<tr>
<td>Unfilled</td>
<td></td>
</tr>
<tr>
<td>Very High</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temporal Demand</th>
<th>How hurried or rushed was the pace of the task?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Low</td>
<td></td>
</tr>
<tr>
<td>Unfilled</td>
<td></td>
</tr>
<tr>
<td>Very High</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance</th>
<th>How successful were you in accomplishing what you were asked to do?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect</td>
<td></td>
</tr>
<tr>
<td>Unfilled</td>
<td></td>
</tr>
<tr>
<td>Failure</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effort</th>
<th>How hard did you have to work to accomplish your level of performance?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Low</td>
<td></td>
</tr>
<tr>
<td>Unfilled</td>
<td></td>
</tr>
<tr>
<td>Very High</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frustration</th>
<th>How insecure, discouraged, irritated, stressed, and annoyed were you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Low</td>
<td></td>
</tr>
<tr>
<td>Unfilled</td>
<td></td>
</tr>
<tr>
<td>Very High</td>
<td></td>
</tr>
</tbody>
</table>
Appendix E – Situation Awareness Rating Technique (SART)

1. Instability of Situation
How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?

<table>
<thead>
<tr>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

2. Complexity of Situation
How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

<table>
<thead>
<tr>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

3. Variability of Situation
How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

<table>
<thead>
<tr>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

4. Arousal
How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

<table>
<thead>
<tr>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
5. **Concentration of Attention**
How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?

- **Low**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7

- **High**

6. **Division of Attention**
How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focused on only one (Low)?

- **Low**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7

- **High**

7. **Spare Mental Capacity**
How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?

- **Low**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7

- **High**

8. **Information Quantity**
How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?

- **Low**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7

- **High**

9. **Familiarity with Situation**
How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?

- **Low**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7

- **High**