# Capability Assessments
## - An Experimental Study of Capability Assessments with Multi-Actor Dependencies

*Malin Hanson & Sebastian Severinsen*

**Division of Risk Management and Societal Safety**
**Lund University, Sweden**

**Riskhantering och samhällssäkerhet**
**Lunds tekniska högskola**
**Lunds universitet**

**Report 5019, Lund 2016**


LUND UNIVERSITY

# Capability Assessments
# - An Experimental Study of Capability Assessments with Multi-Actor Dependencies

**Malin Hanson & Sebastian Severinsen**

**Lund 2016**

Capability Assessments
- *An Experimental Study of Capability Assessments with Multi-Actor Dependencies.*

Förmågebedömningar
- Experimentell studie av förmågebedömningar med beroenden mellan flera aktörer.

Malin Hanson & Sebastian Severinsen

**Report 5019**
**ISSN: 1402-3504**
**ISRN: LUTVDG/TVRH—5019—SE**

Number of pages: 51 (95 including References and Appendices)
Illustrations: Malin Hanson & Sebastian Severinsen

**Keywords**
Capability, Capability Assessments, Risk and vulnerability analyses, Multi-actor, Dependency models, self-assessments, peer-assessments, Mastermind, SPSS, Capacity, Ability, Skill.

**Sökord**
Förmåga, förmågebedömningar, risk- och sårbarhetsanalyser, flera aktörer, beroendemodeller, egenanalys, kamratbedömning, Mastermind, SPSS, kapacitet, skicklighet.

**Abstract**
This thesis investigates how multi-actor dependencies affect the ability to assess capability. It is also investigating if the accuracy of capability assessments for tasks where multi-actor dependencies are present, depends on if capability is assessed individually or with all participating actors. Two multi-actor dependency models were considered representative for all cases. Dependency I represents a scenario where two actors have a common goal, but each actor's individual performance is not affected by the other actor's performance. Dependency II represents a scenario where two actors have a common goal and the first actor's individual performance affect the second actor's performance. It was concluded that the participants tended to be more confident when assessing multi-actor dependency tasks in pairs than individually, although there is no significant difference in accuracy between the assessments. A qualitative analysis suggests that individual capability assessments are more accurate for individual tasks and for Dependency II tasks than for Dependency I tasks. For assessment made with all participating actors there is no significant difference between the dependencies, although assessments for Dependency II tasks tended to be more accurate than Dependency I tasks.



---

Riskhantering och samhällssäkerhet
Lunds tekniska högskola
Lunds universitet
Box 118
221 00 Lund

http://www.risk.lth.se
Telefon: 046 - 222 73 60
Telefax: 046 - 222 46 12

Division of Risk Management and Societal Safety
Faculty of Engineering
Lund University
P.O. Box 118
SE-221 00 Lund
Sweden

http://www.risk.lth.se
Telephone: +46 46 222 73 60
Fax: +46 46 222 46 12

# Preface

This thesis has been written for the Department of Risk Management and Societal Safety at Lund Institute of Technology at Lund University. It has been a frequent, although no longer a fun joke, that we would have to be able to assess our capability to finish this thesis within the given time frame. Despite our masters in risk management, we did not take into account that our experiment would be postponed because of a total University shutdown due to a threat. This was an unfortunate event but showed once again that this is an ever changing world where future events are difficult to foresee.

We would like to give special thanks to:

Henrik Tehler and Hanna Lindbom – our supervisors through this adventure. You believed in what we could achieve and pushed us forward when we needed it. Many thanks for that.

Nils-Erik Eliasson, Anna Hanson, Bo Hanson, Marie Hanson, Emy Hurtig, Kaowe Hurtig, Emilia Severinsen, Eva Severinsen, Leif Severinsen, Eije Olofsson, Evelina Olofsson, Simon Olofsson and Ylva Olofsson – Your comments and participation to the early stages of the experiment were crucial to shape the final result. Without you we would still be stumbling in the dark. Many thanks.

Everyone who chose to take part in our experiment, without you this thesis would not exist.

And to our mascot Lexie, who reminded us that work is no reason to skip lunch and made us stretch our legs every now and then.

Many thanks,

Malin Hanson
Lund 2016

Sebastian Severinsen

## Summary

The modern society is growing fast and is increasingly complex, which inherently increases the number of actors who also are dependent on each other. In order to cater for the increasingly complex society several countries, Sweden and United Kingdom among others, have adopted a capability-based planning approach in their national risk and vulnerability analysis. The capability assessments are conducted in Sweden as part of the risk and vulnerability analyses, which every municipality is required to update annually. However, the term capability and how to conduct assessments of capability have not been defined and structured consistently among practitioners. Since the number of actors who are dependent on each other is increasing, it is crucial to understand how multi-actor dependencies affect capability assessments.

The purpose of this thesis is to investigate how multi-actor dependencies affect capability assessments. It is also part of the purpose to explore if the multi-actor dependencies affect the actual performance depending on the nature of the dependencies. To explore multi-actor dependencies and how to improve capability assessments made under multi-actor circumstances, it is also crucial to understand how the actors perceive the difference between assessing their individual capability and the capability of a group.

There is plenty of solid ground to step on around capability assessments with multi-actor dependencies, with reputable authors in the close-knit fields of risk, vulnerability, capacity and decision-making. However, the particular field of this thesis is in the dark. This thesis aspires to be one of many stepping stones for future research to shed light on capability assessments with multi-actor dependencies.

The foundation of this study is the experiment carried out at two Universities and an upper secondary school in Sweden. A total of 48 participants took part and performed at least one of the two designed tasks. The two tasks were set-up in the game of Mastermind and represented a multi-actor dependency each. Task G1 represents Dependency I, which simulate a dependency where two actors have a common goal, but each actor's individual performance is not affected by the other actor's performance. An individual task, I1, was included in task G1 to be able to derive any differences between individual and group tasks. Task G2 represent Dependency II, which simulate a dependency where two actors have a common goal and the second actor's individual performance depend on the first actor's performance. Before each task, the participants did a capability assessment of their individual and the pair's performance by assessing at which row they would solve the Mastermind code. For each task the participants also conducted a capability assessment together of the pair's performance.

The conclusions are that the participants tended to be more confident when assessing multi-actor dependency tasks in pairs than individually, even so a majority of the participants underestimated both their individual and joint capability to solve the tasks. There was a significant difference between capability assessments and performances for all tasks, including the individual task. This suggests that assessing capability is difficult with and without multi-actor dependencies. However, a qualitative analysis suggests that individual capability assessments are more accurate for individual tasks and Dependency II than for Dependency I. For assessment made in pairs there is no significant difference between the dependencies, although assessments for Dependency II tend to be more accurate than Dependency I.

# Swedish Summary

Det moderna samhället växer fort och ökar i komplexitet. Komplexiteten orsakas delvis på grund av minskat statligt inflytande, vilket följaktligen leder till ökat antal aktörer som i större utsträckning än tidigare också är beroende av varandra. För att kunna ha ett fortsatt bra skydd trots den ökade komplexiteten har flera länder, bland annat Sverige och Storbritannien, börjat med förmågebedömningar i nationella risk- och sårbarhetsanalyser. I Sverige utförs förmågebedömningar som en del utav risk- och sårbarhetsanalyser som kommuner måste uppdatera varje år. Trots detta finns det ingen konsensus kring vad ordet förmåga innebär eller hur förmågebedömningar bör genomföras och vad de bör innehålla. Eftersom antalet aktörer som är beroende av varandra ökar, är det avgörande att förstå hur förmågebedömningar påverkas när flera aktörer är involverade.

Syftet med det här examensarbetet är att undersöka hur beroenden mellan flera aktörer påverkar förmågebedömningar. Det är också syftet att utforska om beroendet mellan aktörerna påverkar själva utförandet med avseende på hur beroendet ser ut. För att undersöka beroendet mellan flera aktörer och hur förmågebedömningar med flera aktörer kan förbättras är det viktigt att förstå hur aktörerna upplever skillnaden mellan att bedöma sin individuella förmåga och förmågan hos en grupp.

Runt ämnet förmågebedömningar med flera aktörer finns det närbesläktade vetenskapliga fält, där exempelvis risk, sårbarhet, kapacitet och beslutsfattning behandlas av väl ansedda författare. Just förmågebedömningar med flera aktörer är dock ett okänt område, där det här examensarbetet siktar på att vara ett första steg mot en tydligare helhetsbild.

Den mest centrala delen av den här studien är experimentet som utfördes på två universitet och en gymnasieskola i Sverige. Totalt deltog 48 deltagare som genomförde minst en av de två uppgifterna. Uppgifterna bestod utav olika utförande av spelet Mastermind och representerade varsin beroendemodell. Uppgift G1 motsvarar Beroende I som representerar fallet där två aktörer har ett gemensamt mål, men där varje aktörs individuella utförande inte påverkas av den andra aktörens utförande. En individuell uppgift, I1, lades till i uppgift G1 för att kunna härleda skillnader mellan individuella uppgifter och gruppuppgifter. Uppgift G2 motsvarar Beroende II som representerar fallet där två aktörer har ett gemensamt mål och där den första aktörens individuella utförande påverkar den andra aktörens utförande. Före varje uppgift genomförde deltagarna en förmågebedömning för deras individuella och för gruppens utförande genom att bedöma vid vilken rad de skulle lösa koden i spelet. Inför varje uppgift gjorde deltagarna dessutom en förmågebedömning tillsammans om gruppens utförande.

Trots att deltagarna tenderade till att vara mer säkra när de bedömde uppgifter där det fanns beroende mellan aktörer i grupp än individuellt, underskattade majoriteten av deltagarna både den individuella och den gemensamma förmågan att lösa uppgifterna. För samtliga uppgifter, inklusive den individuella, fanns en signifikant skillnad mellan deltagarnas förmågebedömningar och utföranden. Det antyder att det är svårt att bedöma förmåga oberoende av vilken typ av beroenden som finns i uppgiften. Dock visar en kvalitativ analys att individuella bedömningar stämde bättre överens med de faktiska utförandena vid individuella och Beroende II uppgifter än där uppgifterna innehåller Beroende I. För bedömningar gjorda i grupp finns det ingen signifikant skillnad mellan de två beroendena, dock tenderar bedömningar för Beroende II stämma bättre överens än för Beroende I.

# Table of Contents

# 1 Introduction

At Lund University the Division of Risk Management and Societal Safety currently runs a project, PRIVAD (Program for Risk and Vulnerability Analysis Development), which among other things develops methods for assessment of crisis management capability. This thesis is part of the PRIVAD project in regard of improving capability assessments.

In this section background, purpose and aim, research questions, restrictions and limitations as well as ethical considerations in relation to the conducted experiment are presented.

## 1.1 Background

The modern society is growing fast and is increasingly complex. One of the reasons is the increased number of actors working together and that these actors are increasingly dependent on each other. The central paradigm of complexity science is a multi-actor system (Heylighen et al., 2007), which in turn can be part of a system of systems. Increased dependencies are what the modern society generates (Lindbom et al., 2015b). The increased number of actors and dependencies are making it particularly difficult to assess capability of an individual system. This because the individual system becomes a fragmented picture of the environment it is part of. To be able to contribute towards a safer and more resilient society it is important to understand these dependencies and incorporate them in risk management.

In order to cater for the modern society several countries, Sweden and United Kingdom among others, have adopted a capability-based planning approach in their national risk and vulnerability analysis (Lindbom et al., 2015a). Previously risk and vulnerability analyses were made with a certain set of scenarios (MSBFS 2010:6). This method is now seen as an inflexible instrument not able to adapt to the new risks that emerge with the increased number of multi-actor dependencies and are therefore subject for recent changes and updates within the Swedish risk and vulnerability analyses.

Capability assessments are conducted as part of the Swedish risk and vulnerability analyses, which every municipality in Sweden is required to update annually. They have been considered not to fulfil the expectations about comparability and user-friendliness (MSB, 2014). The importance of comparability is substantial as the Swedish County Administration Boards are assessing their capability on the basis of the municipalities' capability analyses (Palmqvist et al., 2012). This is one example of a multi-actor dependency. In order to increase the quality and usefulness of the risk and vulnerability analyses in general and capability assessments in particular, the law prescribing the content of risk and vulnerability analyses, MSBFS (2010:6), was replaced by MSBFS (2015:4) and MSBFS (2015:5). MSBFS (2015:5) includes an appendix with indicators for capability assessments in the context of crisis management and emergency preparedness to increase comparability.

The term capability and how to conduct assessments of capability has not been defined and structured consistently among practitioners. Furthermore, capability is commonly used interchangeably with similar words like ability, skill and capacity (Lindbom et al., 2015a). Lindbom et al. (2015a) suggests that capability is the ability to respond depending on capacity, where capacity is defined as the ability to prepare. The concept of capability also has a close relationship with risk, vulnerability and resilience. Therefore, Lindbom et al. (2015a) proposes that the definition of capability is to include the initiating event, the performed task, the consequences of the performed task, the uncertainties concerning the

consequences and the background knowledge. This has to be defined every time capability is used, as capability is not the same in different contexts (Lindbom et al., 2015a). The nature of capability creates new challenges when several actors are introduced, whom are dependent on each other to different extents. These challenges are causing diverse and uncompleted risk and vulnerability analyses on all administrative levels (Palmqvist et al., 2012).

To do assessments can be difficult, especially when what is supposed to be assessed is not properly defined. However, even if capability is properly defined, it is still difficult to assess the actual capability as it is common to overestimate one's capability if there is a history of success and underestimate one's capability if there is a history of failure (Kahneman & Tversky, 1973). It is also common that a response to a complex question is actually based on a simplified version of the complex question (Kahneman & Frederick, 2001). Other psychological aspects like biases and heuristic methods also contribute to uncertainties of capability assessments.

As multi-actor dependencies are increasing due to the complex reality, it is also increasingly important to understand how dependencies affect capability assessments. In this thesis the multi-actor dependencies are divided into two, where the actors are dependent on each other in order to perform and where the actors are not dependent on each other to perform. Both of the multi-actor dependencies are simplified in this thesis in order to be able to make an experiment, but they are in large representative for real situations. An example is the risk and vulnerability analyses performed by Swedish municipalities made separately and therefore not affecting each other, but combined together at a higher administration level where the comparability or lack thereof affects the end product.

There is plenty of solid ground to step on around capability assessments with multi-actor dependencies, with many reputable authors in the close-knit fields of risk, vulnerability, capacity and decision-making. However, the particular field of this thesis is in the dark. This thesis aspires to be one of many stepping stones for future research to shed light on capability assessments with multi-actor dependencies.

## 1.2 Purpose and aim

The purpose of this thesis is to investigate how multi-actor dependencies affect capability assessments. It is also part of the purpose to explore if the multi-actor dependencies affect the actual performance depending on the nature of the dependencies. To explore multi-actor dependencies and how to improve capability assessments made under multi-actor circumstances, it is also crucial to understand how the actors perceive the difference between assessing their individual capability and the capability of a group.

The aim with this thesis is to describe how multi-actor dependencies affect the ability to assess capability. Multi-actor dependencies have not been explored previously in the field of capability assessments. The aim is therefore to add value to the scientific field, this in order to be prepared on how to manage weaknesses and reproduce strengths which appears where several actors are involved.

## 1.3 Research question

The overall research question is,

*How does multi-actor dependencies affect capability assessments?*

The following secondary research questions are important keys in order to create a nuanced response to the general research question:

- Do the capability assessments match the actual performance?
- Is there a difference in accuracy between capability assessments performed individually and in pairs?
- Is there a difference between capability assessments depending on if the assessment was made individually or in pairs?
- Is there a difference between how well the capability assessments match the actual performance depending on if the task was performed individually or in pairs?
- Is there a difference between how well the capability assessments match the actual performance depending on how the multi-actor dependencies were designed?
- Do performances where multi-actor dependencies are present differ from individual performances?

## 1.4 Restrictions and limitations

The following restrictions and limitations apply to this thesis:

- This thesis aims to examine how multi-actor dependencies affect the capability assessments, how multi-actor dependencies affect the result of the designed tasks and perceived difference for participants to perform and assess individually and in pairs. All other areas are considered outside the scope of this thesis
- It is not within this thesis to describe multi-actor dependency properties for other dependencies than those described in the thesis, or for more actors than two.
- The field of implementation is risk- and vulnerability analyses, hence the choice of experimental design.
- In order to draw conclusions, the experiment is a simplification of a complex reality.
- The participants are well aware that they are taking part in a study. They might behave differently if they had encountered the problem outside an experimental setting.
- Group dynamic is crucial where interaction between actors is relevant but is considered outside the scope of this thesis.

## 1.5 Ethical considerations in relation to the conducted experiment

The participants took part in an experiment where the tasks were cognitive. The arrangement took place in surroundings familiar to the participants. If the participants wanted to withdraw their participation, they had opportunities to do so on several occasions. As the circumstances of the experiment were considered safe for the participants' mental and physical well being, a consent form was not distributed to the participants. Instead the participants were informed during the introduction that they were allowed to leave at any time during the experiment.

# 2 Method

In this Section the work process, design process of experiment, research method and statistical methods are described.

## 2.1 Work process

The work process of this thesis is in large divided into six parts. In Figure 1 a flowchart of the work process is shown.


**Figure 1. Flowchart of the work process.**

### *Literature Review*

The aim for the Literature Review was to find relevant literature regarding capability and capability assessments. It was of special interest to find any studies that had been conducted in terms of multi-actor dependencies.

Literature was searched for in the databases LUBsearch, Scopus, Google Scholar and Web of Knowledge. The primary search words were *capabilit*, capability, capacity, capability assessment, capability+risk, self-assessments, self-prediction, group assessments, peer assessments, biases, heuristics.*

### *Define research questions*

The literature within the field of capability assessments in terms of multi-actor dependencies turned out to be scarce. Therefore, the research questions were readjusted to suit the extent of a master's degree thesis and existing literature. The research questions are presented in Section 1.3 and the research method is presented in Section 2.3.

### *Design Experiment*

A solid design of the experiment is crucial to this thesis. The experimental design went through several iterations, including three pilot studies, in order to ensure it measured what was relevant for the thesis. The preparatory work is further explained in Section 4 and the experiment itself is described in Section 5.

### *Conduction of Experiment*

The experiment was carried out on six occasions at four different locations, Lund University, Luleå Technical University, Nils Fredriksson Utbildning and in-house. When the participants arrived, they received a short introduction to the thesis and to Mastermind. They were divided into pairs and had a practice session before they were introduced to the actual task. When the participants had performed one of two tasks, either individually or in pairs, they were given a questionnaire where the participants were able to explain how they had perceived the experiment. A total of 48 participants took part in the experiment.

### *Analyse Results*

In order to analyse the results, statistical testing was made. A software called Statistical Package for the Social Sciences (SPSS) was used for statistical testing. The statistical tests are explained further in Section 2.4 and the results can be found in Section 6.

### *Draw Conclusions*

In addition to the analysed results, conclusions will be made on the basis of the relevant literature as well as the observations made during the experiment. The discussion can be found in Section 7 and the conclusions in Section 8.

## 2.2 Developing design features

The design process of the experiment has been an iterative process, where proposed activities have been reviewed several times as the design features have developed. To suit the purpose of the thesis, the activity had to meet a set of Design Features, such as time limits and activity criteria. During the iterative process the design features were developed and rejected continuously. The following flow chart shows a schematic overview of the process.

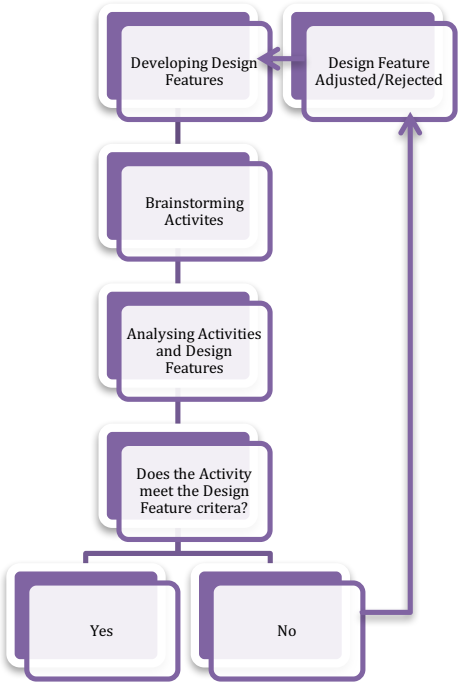



**Figure 2. Flowchart of the iterative process of developing design features.**

## 2.3 Research method

There are a number of different research methods available when conducting research, each suited for different purposes. They provide a general description of how a researcher chose to work with their project. The chosen research method is not fixed throughout a project but different parts of a project might benefit from different methods. To describe the main research method serves to support the authors through deciding the aim and purpose of their project towards increased knowledge within the field of research. The main purpose of a study might be either descriptive, exploratory, explanatory or problem solving (Höst et al., 2011).

For this thesis a mix between a descriptive and explanatory study with an experimental design is considered to provide the best explanation. A descriptive study serves to describe how something works, while an explanatory study serves to find causation and explanations as to how the studied phenomena works (Höst et al., 2011). The primary data will be quantitative and come from observations and results during conduction of the experiment. Secondary data will be partly quantitative but mainly qualitative and come from a questionnaire each participant will take part of. Where replies to the questionnaire are considered to require clarification, the authors will conduct short follow-up interviews when the questionnaires are returned. The interviews are only to clarify the intent of the questionnaires replies.

Due to the choice of research method a fixed design is used throughout the study. This means that nothing can change when the live experiment has begun, even if changes would improve the experiment. All issues regarding the experimental design and the questionnaire are required to be identified and dealt with before the live experiment, as a change in design would create a discrepancy. Discrepancies in the circumstances of which each participant conducts the experiment would affect the comparability of results. This makes the design process including pilot studies an important part of this thesis.

The questionnaire is designed with both open and closed questions and contains both quantitative and qualitative data. The main purpose of the questionnaire is to provide information about the experiment from the participants' perspective.

The sample selection is treated with a quasi-experimental design, which means that the participants are not randomly selected from the population. Instead, the sample was based on availability among the population. However, non of the participants were assigned a specific task with any regards to personal properties or traits. The aim with the sample selection was to resemble the population who are involved with risk and vulnerability analyses. I.e. participants with various backgrounds such as social science and natural science.

## 2.4 Statistical methods

For the purpose of managing quantitative data from the experiment Statistical Package for the Social Sciences (SPSS) was used. SPSS is a program suitable for statistical analysis for a wide range of different kind of data. Simple descriptive statistics for qualitative data is also conducted with SPSS.

Hypothesis testing is conducted through comparison of means which in SPSS is achieved through T-tests.

T-test belongs to the parametric methods of statistics which provide better precision in the analysis then non-parametric methods. The advantage is that the parametric methods are more sensitive towards finding a statistical difference in the available data.

The following assumptions should be fulfilled for maximum accuracy of a T-test. Adapted from Cunningham and Wallraven (2012, pp.260-81) and Körner and Wahlgren (2006, pp.182-233).

- *Equal sample sizes.* Unequal sample sizes require a correction to the t-statistics. This is computed automatically in SPSS and only applies to independent samples T-test.

- *Equality of variance.* Unequal variances require correction to the degrees of freedom and affect the power of the T-test. SPSS automatically conducts Levene's test to determine equality of variance.

- *Normal distribution.* The T-test is able to handle some difference in the data distribution, although heavily skewed data might prove a problem. The alternative is to use a non-parametric method which do not require the data to follow a normal distribution.

- *Repetitions and multiple measurement categories.* T-tests can not compute multiple repetitions from the same sample, i.e. the participants can not repeat the experiment multiple times and count towards a better result. The alternative is to use a paired sample T-test. However, the same restriction applies, only one pair of data is comparable at a time. This kind of design requires different methods of analysis.

- *Multiple T-tests.* To conduct multiple test on the same set of data increases the chance to get a Type 1 error. In other words, there is a greater chance to find a significant difference between two groups even though in reality such difference does not exist.

**Independent samples T-test**
Compares the mean of two independent groups on the same dependent variable (Cunningham & Wallraven, 2012, pp.260-63). SPSS produce descriptive statistics like mean, standard deviation and frequency as a basis for conducting the test. The result consists of a t-value, degrees of freedom, significance level and a confidence interval. In addition, SPSS also conducts Levene's test to validate the assumption of equal variance.

**Paired sample T-test**
Is used for dependent measurements. For example, when the same participants are part of both groups that will be compared, i.e. they are tested twice. Paired sample T-test is applicable when the experiment uses within-participant design and have more statistical power then the independent sample T-test. The benefit is that the variances of the groups have less effect on the results which in turn means that the sample sizes can be smaller. This is achieved through not calculating the mean and variance of each group separately but from the difference between the paired samples (Cunningham & Wallraven, 2012, pp.279-81).

**Levene's test for equality of variance**
Tests the hypothesis that the variance in two groups are equal. The main purpose of this test is to test the assumption of equal variance which accompanies T-tests for example. Levene's test shows significance when there is a discrepancy in the variance and might suggest a different test is required for this specific set of data. SPSS deals with this automatically in relation to T-tests.

**Cohen's d for effect size**
Measures effect size when comparing means, independently from the variables but dependent on which statistical test is used. The purpose of this test is to determine if the results are relevant through establishing the effect size. An example to illustrate the concept of effect size could be the following situation; the results from a statistical test shows a "small but significant" correlation between two variables (Cunningham & Wallraven, 2012). This means that the difference found in the samples is significant but says nothing about the relevance of

the claim. Effect size is a measure that compliments the probability value by indicating the size of the effect and therefore provides a more holistic view of the results.

Cohen's d is used to support the results from the T-tests but will not be central to this thesis. Calculations of Cohen's d are only possible to do after completion of the experiment and the statistical analysis which is why the results do not affect any design features or other properties within this thesis even if the results suggest a small effect size.

# 3 Literature study

A thesis is to add value to previous studies and finds. In order to do so, a literature study is crucial to find out what already have been discovered and discussed. An efficient literature study is systematic and well documented (Höst et al., 2011). Kahn et al. (2003) proposes a literature study methodology of five steps to be used iteratively:

1. Framing questions for review
2. Identifying relevant work
3. Assessing the quality of studies
4. Summering the findings
5. Interpreting the findings

The chosen review questions are as follows:

- What is the definition of capability?
- Where and when are capability assessments used?
- How is capability assessed and how good are the assessments?
- Why are multi-actor dependencies within capability assessments interesting?

## 3.1 What is the definition of capability?

There is no agreed definition of the word capability and it is often used interchangeably with capacity, ability and skill (Lindbom et al., 2015a). One of the reasons for a non-agreed definition could be that it is a commonly used word. This means that people associates a meaning to the word. While the intuitive meaning might not be wrong, it often comes with a feeling that one understands the word and assume others have the same intuitive meaning to the word (Dekker & Hollnagel, 2004). When a word has been used for some time, the "Emperor with no clothes" phenomena may appear according to Dekker and Hollnagel (2004). This means an involuntary consensus has been made that no one asks for the intent of the word since everyone already understand the intent of the word (Dekker & Hollnagel, 2004). This can be challenging since the interpretation of capability is different in different contexts. For example:

- "Capability is the characteristics, abilities and resources that facilitate a specific, or predictable response from and within the various elements of the critical national infrastructure" in regards of finding vulnerabilities within critical national infrastructure (Hills, 2005, p.13).

- "Capability is the resources, systems, structures and processes necessary to deliver (currently and in the future) the required level of performance in fulfilment of the mandated objectives" in regard of stability of business (Bhatta, 2003, p.403)

- "Capability is the ability and capacity to attack a target and cause adverse effects" in terms of intelligent analysis of counter terrorism (Haimes, 2006, p.293).

Lindbom et al. (2015a) chose to include definitions of risk, vulnerability and resilience into their definition of capability. The definitions of risk, vulnerability and resilience used for this purpose are as presented by Aven (2011):

Risk = (A,C,U), the uncertainty (U) about and severity of the consequences (C) of an Activity (A).
Vulnerability = (C, U | A), the uncertainty (U) about and severity of the consequences (C) of the activity given the occurrence of the initiating event *A*.
Resilience = (C, U | any A), the uncertainty (U) about and severity of the consequences (C) of the activity given the occurrence of any type of activities (A).

Lindbom et al. (2015a) therefore suggest the following definition for capability:

Capability = ($C_T$, U | A, T), the uncertainty (U) about and the severity of the consequences (C) of the activity given the occurrence of the initiating event (A) and the performed task (T).

In order to assess capability, and not incapability, the consequences are described as the positive outcome (Lindbom et al., 2015a). This is the opposite of the more commonly described consequences for risk, vulnerability and resilience.

## 3.2 When and where are capability assessments used?

Capability assessments is an interdisciplinary activity that can be found on all levels of society, from an individual to a global level. This requires the methods used for conducting capability assessments to be flexible and adapt to different circumstances (Palmqvist et al., 2014).

Capability assessments on an individual level are sometimes referred to as self assessments or self predictions and are often related to different kinds of tests where experts within the field provide the benchmark for a correct assessment. Sung et al. (2010) present a study where seventh graders in a musical class assess their own performance when using a recorder. Their self assessments were compared with the teachers' assessments of their performance. Another example is Mynttinen et al. (2009) who studied novice drivers' self assessed driver competence compared to assessments made by driving examiners. The act of self assessment is difficult to do without being affected by numerous psychological phenomena that are working against achieving an accurate objective result (Dunning et al., 2004). These problems do not only affect individuals but they affect businesses, and organizations as well. However, Dunning et al (2004) continues by stating that organizations can through routines and procedures reduce the impact of these self-assessment biases.

Capability assessments on an organizational level may vary in complexity. Palmqvist et al. (2014) found through a scoping study with focus on risk and crisis management, that a common purpose for conducting capability assessments was as a basis for decision making. Assessing the possibility to complete a project within a fixed set of parameters, such as completion time, budget and resources suggest that a number of different disciplines are involved e.g. economy, engineering and management. Organizational capabilities are often derived from its human resources, their employees' skills and abilities. However, Bhatta (2003) suggest that taking a more holistic perspective might serve to develop organizational capabilities to not only include human resources.

In Sweden, municipalities, county councils and governmental agencies are required by law to conduct capability assessments (SFS 2006:544) (SFS 2006:942). These are comprehensive assessments which focus on crisis management and emergency preparedness. They are encompassing many different social sectors e.g. healthcare, communication, infrastructure, resources, and information. This kind of capability assessment is a good example of why it is important to consider multi-actor dependencies. Palmqvist et al. (2012) describes how municipalities, county administration boards and governmental agencies to a varying degree involves external actors. The involved organizations sometimes assess their own capability while sometimes assessing the capability of the geographical area or a specific social sector.

The answer to the question, when and where capability assessment is used, is closely linked to which definition one apply. When forced to resort to an intuitive understanding of the term capability, the variations of applications and interpretations are endless (Palmqvist et al., 2012) and suggests that the answer to the question is, always and everywhere.

## 3.3 How is capability assessed and how good are the assessments?

Capability is assessed by everyone everyday as in terms of catching the bus on time or how much work will be done before lunch. As it would take to much energy to think about all choices and all interpretations being made, the human mind take shortcuts. The shortcuts are called heuristics and affect the outcome when making a decision or assessment (Kahneman, 2011, p.75). Kahneman (2011, pp.105, 92, 75 ) points out the following among others:

**Availability heuristic** includes that the relative importance between subjects are dependent on how "available" they are. For example, the possibility of dying in an accident is rated much higher than dying in a disease, even if the second is a lot more common than the first. This heuristic could affect a capability assessment for example by overestimating the importance of a factor/task/asset or underestimating the importance of a factor/task/asset depending on if it has been previously discussed or not.

**Anchor heuristic** was considered proved when Kahneman (2011, p.92) made an experiment showing that when presented to a specific number, this number affected the answer to questions where it was irrelevant. For example, participants were given a certain number by a wheel of fortune, 10 or 65, and were later asked which fraction they estimated African countries made of the total number of countries in United Nation. Those who had been exposed to 10 estimated a lower fraction than those exposed to 65.

**Attribute substitution** is usually triggered when complex questions are asked. It tends to make people to answer an easier version of the question. For example, if the question is "how happy are you with your life?" the question usually being answered is "how happy are you right now?".

Heuristics cause people to make mistakes, even experts. Expertise is about reacting to cues (Kahneman & Klein, 2009) and when an expert reacts to one or several cues, a solution arises intuitively (Klein & Calderwood, 1991). The solution is either modified to suit the purpose or rejected. If it's rejected, another solution is analysed until a suitable solution is found (Klein & Calderwood, 1991). This is called the Recognition-Primed Decision model (RPD model) and is considered to be a more correct model for decision-making by experts than the traditional tree model (Klein, 2008). The tree decision model is used when several solutions are analysed throughout to find the best solution, which are more common for novices (Klein

& Calderwood, 1991). An example in terms of capability assessments is the result of the survey by Lindbom et al. (2015b) which suggests that more experienced within the fire rescue service are able to make a more accurate capability assessment with less information. I.e. an experienced rescue service official can describe capability as a tank truck, since the experienced within the fire rescue service know what potential a tank truck has to a greater extent then a beginner (Lindbom et al., 2015b).

Kahneman and Klein (2009) argue that although experts' intuitive solutions usually are applicable, they are not always suitable for a specific task. However, they found that "true experts" tend to know what they do not know, while non-experts do not know what they do not know. Skilled professionals, such as judges and fire fighters, are usually unaware of which cues that affect their intuitions, but non-skilled people are even less aware. Even the most unskilled person could have success once, which Kahneman and Klein (2009) means is often reoccurring on the financial market. Kahneman (2011, p.19) mentions cognitive illusions which occur where decisions are being made on routine. The best way, according to Kahneman (2011, p.19) to prevent mistakes following cognitive illusions is to learn under which circumstances mistakes are being made and being cautious in these situations.

How well capability assessments correspond to the actual outcome have been tested in different forms, although often called self-prediction, self-assessment and peer-review among others. In general, the studies contain one or more of the following properties; individual self-assessment, group self-assessment, individual peer-assessment and group peer-assessment. Table 1 is a summary of the relevant articles and their properties.

**Table 1. Summary of properties of literature which included at least one of the following categories: individual self-prediction, group self-prediction, individual peer-prediction and group peer-prediction.**

| Author | Individual self-assessment | Group self-assessment | Individual peer-assessment | Group peer-assessment |
|---|---|---|---|---|
| **Halkjelsvik et al. (2010)** | x | | | |
| **Mynttinen et al. (2009)** | x | | | |
| **Balcetis et al. (2008)** | x | | x | |
| **Dunning et al. (2004)** | x | | x | |
| **Fredriksson et al. (2011)** | x | | x | |
| **Sundvik and Lindeman (1998)** | x | | x | |
| **Vallone et al. (1990)** | x | | x | |
| **Sung et al. (2010)** | x | x | x | x |

- Halkjelsvik et al. (2011) showed that people in general underestimate the timeframe of large tasks and overestimate the timeframe for small tasks. Coherently, people tend to estimate that they are able to do more work per time unit for larger tasks than for small tasks.

- Mynttinen et al. (2009) made drivers from Finland and the Netherland assess their driving skills in different areas, for example vehicle control, recognising and avoiding risks. The participants' assessments were higher in both samples than the examiners' assessments. It was concluded that 40% of the Finns and 50% of the Dutch made realistic assessments of their skills.

- Balcetis et al. (2008) investigated how accurate people assessed positive and negative behaviours of themselves and their peers. They concluded that individualistic cultures, i.e. many western countries, overestimated their generous manner and underestimated their negative behaviours, although they were about right regarding their peers. Members of collective cultures on the other hand had more accurate prediction both regarding their own and their peers' positive and negative behaviour (Balcetis et al., 2008).

- Dunning et al (2004) describes in their literature summary that people in general state that their skills in ambiguous traits, such as being sophisticated, idealistic, and easier tasks, such as riding a bike, are above average. Less ambiguous traits, such as being neat and athletic, people in general do not consider themselves being above average. Further a majority state that they are more likely than their peers to do a correct self-assessment. Dunning et al (2004) also describes that people underestimate the time it will take to complete a task. For example, students where asked to estimate how certain they were to meet a set of deadlines. For all of them, the confidence far exceeded their achievements.

- Fredriksson et al. (2011) used self-assessments to assess pupils' self-perception of academic ability. About one quarter of the students made accurate assessments of their reading ability, while the majority over-estimated their reading ability. The older students seemed to be more accurate in their assessments, although the difference was not significant. Fredriksson et al. (2011) assumed the gap between the groups would be larger, but explains the small gap with the older students' inability to self-reflect as they take their reading ability for granted.

- Sundvik and Lindeman (1998) asked a sample of salespeople to assess their individual productivity. The individual assessment was compared with assessments made by a supervisor who was familiar with the salesperson and a supervisor who wasn't familiar with the salesperson. The highest assessment was made by the salesperson, the supervisor familiar with the salesperson almost put as high scores as the salesperson. The supervisor unfamiliar with the salesperson put the lowest score. However, it was not part of the study how well either assessment agree with reality.

- Vallone et al. (1990) let college students answer 41 Yes or No questions about events that may happen, for example "hours study per day >2", "roommate become best friend" and "plan post-graduate degree", and how certain they were it would occur for themselves and their roommate. The participants were constantly overconfident in their predictions although the study showed that they were more confident and

accurate in their self-predictions than peer-predictions. However, as confidence increased, the gap between accuracy and confidence widened.

- Sung et al. (2010) asked adolescents to assess their individual performance where the task was to perform musically or to create a website. The participants were also divided into groups to solve similar tasks. While in groups they together assessed the performance of the group as well as the performance of other groups. In general, the high-achieving groups underestimated their performance, both individually and in groups, while low-achieving groups overestimated their performance, both individually and in groups. The peer-assessments of other groups were lower than the teachers' assessments for all groups, but the difference between the peers' assessment and the teachers' assessment of a high-achieving group was more distinct.

## 3.4 Why are multi-actor dependencies within capability assessments relevant?

Results from a number of studies regarding capability assessments have shown the need to include multi-actor dependencies when assessing capability. The study by Palmqvist et al. (2012) originates in developmental work regarding capability assessments in relation to Swedish risk and vulnerability analysis. They found that capability assessments should consider dependencies between different actors and systems. Palmqvist et al. (2012) motivates this by pointing out that assumptions about the environment in which capability is assessed will affect the outcome. For example, if it is assumed that the power supply will function normally or not will have an effect on the capability to deal with an emergency or disruption of some sort. Since these dependencies between actors exist, it is important to include this information in a capability assessment Palmqvist et al. (2012). Furthermore, there is a possibility that without considering dependencies the assumptions will lean towards some kind of normality when assessing capability. In other words, the assumption is that your organization is the only one affected by the disrupting event. This affects, not only the specific actor's capability but might also cause discrepancies between the assessed and actual capability when conducting joint assessments (Palmqvist et al., 2012). Another note from Palmqvist et al. (2012) is the connection between capability assessments and critical dependencies. Both activities appear within risk and vulnerability analysis however, they are rarely combined in a way where the critical dependencies are affecting the outcome of a capability assessment. If this connection is recognized by the actors there is much to gain within both activities and the same goes for other activities, a holistic perspective is to strive for when dealing with emergencies and crisis management.

Participants from a study conducted by Lindbom et al. (2015b) provided their opinions about the usefulness of capability assessments for decision making. They mentioned collaboration partners as something important to include in the resource description and how a multi-actor perspective affects how each actor communicates their own capability. A pragmatic example of the issues might be if a fire and rescue service only describes their capability by listing how many fire fighters and fire trucks they have. A dependent actor, a hospital for example, who are not familiar with the fire and rescue service capabilities are not able to figure out how this affects their capability. Lindbom et al. (2015b) states that the need for multi-actor capability assessments is sprung from the complexity of today's society. Which in turn demands that the tools used are adapted to the current circumstances.

Lindbom et al. (2015a) propose a definition of capability which would make it easier to relate to concepts like risk, vulnerability and resilience. To use a definition that is accepted and approved by those who work in close relation to the specific concept is an advantage when it comes to developing common practice and all kinds of developmental work. Considerations regarding multi-actor dependencies were taken into account throughout this process and Lindbom et al. (2015a) stresses that if joint capability assessments are to be successful it is important that each actor's capability description is compatible with others. The need could come from an example as the rescue service cannot assess their ability to put out a fire in an area where they need escort by police if they neglect to take into account the police capability to provide escort. The definition provided by Lindbom et al. (2015a) facilitates functions needed when conducting multi-actor assessments, they state that capability based planning gains in popularity and capability assessments need to follow the developmental stages of society. That includes incorporating multi-actor dependencies in capability assessments.

## 3.5 Summary

There is no agreed definition of capability which makes it hard to achieve comparable assessments. Since capability is often used interchangeably with capacity, ability and skill, this generates an intuitive meaning to the word. Within the field of risk there is an effort to define capability and make capability assessment easily applicable to other activities such as risk assessments. This is done within the same framework as risk, vulnerability and resilience.

The methods used to assess capability are just as diverse as the definitions of capability. Subconsciously the human mind takes shortcuts to lessen the burden of too many and complex alternatives. This is called heuristics and plays a huge part in decision-making processes. Experts are generally better at keeping their subconscious self at bay when conducting assessments, one reason is because they know what they do not know and can therefore choose to ignore the subconscious shortcuts.

Capability assessments are frequently conducted to serve as a basis for decision-making and can be found everywhere in society from an individual to a national level. This is not exclusive for the field of risk management but relevant for all sectors where capability assessments are used. The most common observations from previous studies are that the participants in general overestimate their capability (Mynttinen et al., 2009) (Dunning et al., 2004) (Fredriksson et al., 2011) (Vallone et al., 1990) and are overconfident in their assessments (Vallone et al., 1990) (Sung et al., 2010) (Fredriksson et al., 2011).

Joint capability assessments are already implemented. For example, the Swedish County Administration Boards' evaluating and compiling the municipalities' risk and vulnerability analyses. In this case the Swedish County Administration Boards are dependent on each municipality to provide a comparable capability assessment. However, since the society grows more complex, partly caused by a decreased governmental influence and an increased number of actors, it creates challenges to understand the vulnerabilities of a system. To avoid future surprises, joint capability assessments among concerned actors are required to increase cooperation around established responsibilities.

# 4 Preparatory work

Once the research questions for this thesis were clearly defined, it was apparent that an experiment was required as there were no relevant literature available within the field the authors wished to study. The idea was to measure how well people are able to assess their capability given a specific task with multi-actor dependencies. Finding a specific task, simple enough to avoid unnecessary unknown parameters affecting the outcome but yet representative for capability assessments made in risk- and vulnerability analyses, turned out to be rather challenging. The challenge to create an appropriate experiment to suit the conditions inherently required by the research question was an iterative process, which is described in this Section.

## 4.1 Design features

During the process of developing a suitable experiment, there have been certain features crucial for the experimental design. In Section 4.1.1-4.1.3, three of the most important features are presented. The features have been developed with the process described in Section 2.2. In addition to those, the following features were considered important to implement:

- *Possible to measure performance.* A straightforward way of measuring performance is crucial to reduce time spent on administration before, during and after the experiment.

- *Simple instructions.* The task is required to be simple to explain in order to minimize misunderstandings and reduce time spent on administration.

- *Cognitive task.* A cognitive task is considered more valid than a time perception task or a physical task.

- *Possible to conduct indoors.* The experiment environment and surroundings are easier to control, and participants are considered more likely to attend if the experiment is conducted indoors.

### 4.1.1 Dependency models

In order to design an experiment for multi-actor dependencies, the authors created a dependency model for this specific case where capability is the critical parameter. The model is simplified in order to be able to cater for multiple scenarios and cater for two actors, although the same principles apply for cases with more actors.

The two actors represent different stakeholders whom take part in a capability assessment. Actors may be individuals, groups of people, departments within a company, companies, organisations, and administrative authorities. The actors are considered to have the following characteristics and properties;

- Responsible for different parts of a task.
- A common goal
- Limited resources

Resources may be:

- Knowledge/Competencies
- Staff
- Equipment (for example vehicles and software)
- Material (for example sandbags and gurneys)
- Physical capability
- Time

The two actors are dependent on each other in order to reach the common goal. The dependences are either;

Dependency I: The performance of each actor is essential, but a poor performance of actor 1 does not affect the performance of actor 2.

Example: A situation where two actors are dependent on each other to reach a common goal, for example running relay. If actor 1 underperforms it affects the result, but it does not affect how fast actor 2 are able to run.

Dependency II: The performance of each actor is essential and poor performance of actor 1 affect the performance of actor 2.

Example: A scenario where two actors are dependent on each other and have a common goal is when a rescue service and a hospital are working together to save a burn victim. The hospital treating the burn victim has limited resources of staff, knowledge, equipment and time. The longer the victim is exposed to fire, the more resources are required to save the patient. At a certain point the resources are not enough and the goal is not reached. This means the hospital is dependent on a fast response and rescue from the rescue services in order to save a burn victim.

### 4.1.2 Participants

Initially the aim was to recruit fire engineering and risk management students only since they have a natural connection to crisis management which is the focus of risk and vulnerability analyses. Due to recruitment issues, it was required to enlarge the sample to include participants from other fields of studies. In addition to engineers, social workers and medical personal also take part in risk and vulnerability analyses. Therefore, the wider spectrum is considered to be more representative than the original idea. Section A.1 in Appendix A presents detailed information regarding the properties of the participants.

### 4.1.3 Time limitation

To make the experiment accessible and inviting to the chosen population, the schedule followed the university standard, i.e. starting at 15 min past full hour and lasted for 45 minutes. By using this approach, the potential participants could participate when they had a longer break between lectures without clashes to their on-going schedule.

### 4.1.4 Choice of experimental activity

Activities that have been considered during the design process are presented below with their proficiencies and constraints.

*Running relay*
Two actors estimate their own and joint capability to run a certain distance to meet a certain time.

Proficiencies:

- Easy to measure performance.
- It is possible to find two distinct groups (beginners and experts). Athletes from the local track and field club as experts and general public as beginners.

Constraints:

- It is physically challenging for the participants.
- Participants might choose a safe time and aim for precision instead of maximum effort.
- It is neither a complex nor a cognitive activity.

*Standing long jump*
Two actors estimate their own and joint capability to jump a certain distance.

Proficiencies:

- Similar to the running relay experiment but not as physically challenging.
- Easy to measure the performance.
- Can be conducted indoors
- It is neither a complex nor a cognitive activity.

Constraints:

- Hard to know if there is an expert group within this field.
- Participants might choose a safe length and aim for precision instead of maximum effort.

*Candy Crush*
Two actors estimate their own and joint capability to solve a game.

Proficiencies:

- Possible to control each participant's resources.
- Can be conducted indoors.
- Participants are likely to have previous experience of Candy Crush.

Constraints:

- Hard to find a group of experts.
- Hard to measure performance.
- Involve several technical gadgets like phones or tablets.

*Exams*
Two actors estimate their own and joint capability to perform at an exam.

Proficiencies:
- A suitable complex and cognitive activity.
- Easy to measure the performance.
- High accessibility since the activity already exist.

Constraints:
- Ethical considerations regarding assessing peers.
- May affect the results of the exam subconsciously.

*Puzzle*
Two actors estimate their own and joint capability to solve a puzzle.

Proficiencies:
- Possible to control each participant's resources.
- Can be conducted indoors.

Constraints:
- Hard to distinguish different levels of expertise.

*Time management*
Two actors estimate their capability to tell when a certain time is met, for example to estimate when 5 min in total has passed.

Proficiencies:
- Possible to cater for different situations.
- Can be conducted indoors.

Constraints:
- Hard to measure relevant performance.
- Hard to distinguish different levels of expertise.

*Mastermind*
Two actors estimate their own and joint capability to solve the code, i.e. at which row they will solve the code at.

Proficiencies and constraints for Mastermind are presented in Section 4.2.2 as Mastermind is the task used for the experiment.

## 4.2 Mastermind
The game called Mastermind is played on a game board, the game board is shown in Figure 3 and in Appendix D. To win the game, a code of four dots is to be solved. Each dot can be one out of six colours. All colours can be used 0-4 times, which means a code can be of a single colour, four different colours and everything in between. When each row is filled, feedback is given. A black feedback dot means one dot is the right colour in the right position, a feedback cross means one dot is the right colour but in the wrong position, and a white feedback dot means wrong colour. The order of the feedback is black, cross and white as the feedback dots

do not represent a specific ”guess dot”. After a couple of rows conclusions can be drawn of which colours are in the code and where they should be located.

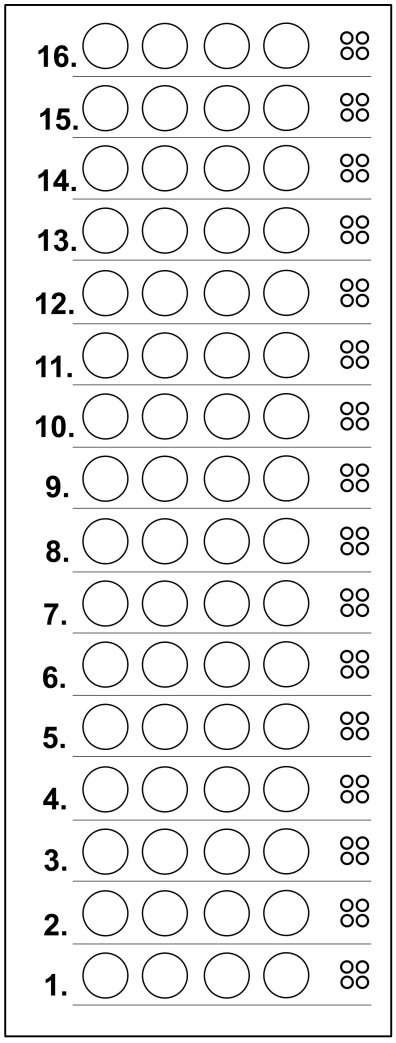



**Figure 3. The game board used for the experiment. The big dots being the "guess dots" and the smaller dots being the "feedback dots".**

### 4.2.1 Initial task design

During the early experimental design several tasks were discussed. The tasks were based on Mastermind in different settings, where the participants were assessing their capability through assessing at which row they would solve the code. The proficiencies and constraints in Section 4.2.2 and the pilot studies described in Section 4.4 are based on the following tasks:

**I1** - The participants solve a code each, independent of each other.
**G1** - The participants solve a code each and the result is combined together. This is designed to represent Dependency I.
**G2** - The participants solve a code together, one participant starts and the other finishes. This task is designed to represent Dependency II.

The final experimental design is described in Section 5.

### 4.2.2 Proficiencies and constraints

Mastermind inherently includes several important parts of the experiment. The following proficiencies are considered important:

1. It’s a cognitive task
2. The majority of the participants will be students of engineering, i.e. used to think logical and understand instructions quickly.
3. It is possible to change the resources, e.g. how many dots the code is, the amount of rows the participants are allowed to use, if each colour is or is not repeated
4. Dependency II can be simulated. The first participant waste resources (rows) if she/he is inattentive. It is critical the second participant understand how the first participant was thinking in order to gain progress from used rows.
5. The participants have a clear common goal.

Even if Mastermind inherently includes several important design features of the experiment, the following constraints are required to be considered:

1. It takes time to solve a game for inexperienced participants, which could lead to fewer data points.
2. The participants are likely to have limited experience of the game.
3. There is only one solution to each game.

4. There is an element of uncertainty in the game, which means that participants' first guess could be correct.
5. The game is usually solved between rows 5-7 when experienced. This is based on the authors' experience through preparatory work including pilot studies.

To minimize the impact on the result, the constrains were managed as follows:

1. The number of tasks was limited to suit the given time frame.
2. The participants were given time to practice within the experiment, before the tasks were explained and conducted.
3. There is only one solution to each game, but there are several ways to find it, therefore this is not managed further.
4. To control the level of uncertainty, all randomized codes used during the experiment consisted of four different colours. I.e. there where no codes with two, three or four of the same colour. A set of six codes was developed for each supervisor.
5. By restricting the amount of rows available for the participants where dependency II is tested.

## 4.3 Definition of capability

Since there is no consensus regarding the definition of the word capability it is required to be defined where it is used. In the context of this thesis, the definition of capability is the ability to perform a specific task and reach a certain result. The participants who assessed their capability to solve the code were therefore asked to assess at which row they would solve the code.

## 4.4 Pilot studies

Before the experiment was carried out, three pilot studies were conducted. Each study resulted in changes to the experiment itself and the material used during the experiment. The pilot studies were a crucial part of the preparatory work leading to the experiment. In this section the pilot studies are described.

The first and second pilot study consisted of four tasks, two individual tasks and two group tasks as well as time for practice and a questionnaire. Results from the first pilot study suggested that in order to complete the entire experiment, the participants required roughly twice the allocated time of 45 minutes. It was apparent that the participants, who in general had no previous knowledge of Mastermind, needed more time to complete the tasks than anticipated.

However, a few latent problems were recognized as the reason behind the huge discrepancy between the suggested amount time and the required time during the first pilot study. One of these issues was that the participants spent a lot of time contemplating their first two or three guesses. Where, in reality, there is little information to take into account. This was in other words a tactical error that exposed inexperience and uncertainty among the participants. Consequently, the authors decided to put a time limit on completing each task in the experiment with the aim that this would encourage the participants to manage their time better.

Another issue that surfaced during the first two pilot studies was related to the design of the group task G2, which represent dependency II. In short the task G2 involves two participants, where the second is dependent on the first to solve the code. G2 is described in full in Section 5.3.2. The results showed that the person who completed the first five guesses could manage to solve the code before the switch had taken place. This completely removed the aspect of the participants being dependent on each other, i.e. it no longer fulfilled the purpose for this study.

The third pilot study was streamlined and consisted of a practice session and two group tasks. The participants did the questionnaire too, but at this time the questionnaires were to be sent to the participants via email in order to minimize administration time within the limited timeframe of 45 minutes. The two individual tasks had been removed, as there was no time to investigate how important experience was in order to assess capability. However, individual data could still be gathered from the first group task (G1, see Section 5.3.1) as the participants solved one code each.

As a time limit would decrease the participants' resources to solve the code, it was decided that there would be no time limit for each task. However, a recommendation was made to the participants to not spend too much time between row 1-4 and solve the code in a steady pace instead of as soon as possible. This was to avoid the early 'clogging' discovered in the previous pilot studies.

There were two important finds during the third and last pilot study. First it still took too long to solve the codes and second that the questionnaire could be hard to understand. Despite the authors' effort to make it easier to understand from previous pilot studies, it still required to be explained. It was therefore decided that the experimental design was to include a practice session, one of the two group tasks and the questionnaire.

The final experimental design is presented in full in Section 5.

## 4.5 Research questions and hypotheses

To be able to answer the research questions, a set of hypotheses were created for statistical purposes.

As several research questions require more than one hypothesis, the hypotheses do not correspond to a specific research question. Instead hypotheses, whole or in part, may be used for more than one research question. Some hypotheses are task specific where I1 represent the individual task, G1 the task performed in pairs representing Dependency I, and G2 the task performed in pairs representing Dependency II.

The relationship between research questions and the hypotheses are as follows:

- Do the capability assessments match the actual performance?

  **Hypothesis 1:** There is no difference between capability assessments and performances.

- Is there a difference in accuracy between capability assessments performed individually and in pairs?

  **Hypothesis 2:** There is no difference in accuracy between capability assessments made individually and in pairs for G1.
  **Hypothesis 3:** There is no difference in accuracy between capability assessments made individually and in pairs for G2.

- Is there a difference between capability assessments depending on if the assessment was made individually or in pairs?

  **Hypothesis 6:** There is no difference between capability assessments made individually and in pairs for G1.
  **Hypothesis 7:** There is no difference between capability assessments made individually and in pairs for G2.

- Is there a difference between how well the capability assessments match the actual performance depending on if the task was performed individually or in pairs?

  **Hypothesis 4:** There is no difference in accuracy for individual assessments between tasks (I1, G1, G2).

- Is there a difference between how well the capability assessments match the actual performance depending on how the multi-actor dependencies were designed?

  **Hypothesis 4:** There is no difference in accuracy for individual assessments between tasks (I1, G1, G2).
  **Hypothesis 5:** There is no difference in accuracy for pair assessments between tasks (G1, G2).

- Do performances where multi-actor dependencies are present differ from individual performances?

  **Hypothesis 8:** There is no difference between performances of task I1 and G2.

# 5 Experiment

In this chapter the practical parts of the experiment are described.

## 5.1 Locations

The experiment was carried out on six occasions:

- Twice at Lund University
- Once at Nils Fredriksson Utbildning (upper secondary school)
- Once at Luleå Technical University
- Twice in-house.

It was conducted in small rooms, separate from other activities within the building.

## 5.2 Participants

The participants from Lund University (LU) and Luleå Technical University (LTU) were recruited through an email sent to Fire Engineering students (LU and LTU), students of Risk Management and Safety Engineering (LU), and students of Disaster Risk Management and Climate Change Adaptation (LU). The word of mouth spread the experiment further and increased the number of participants. As the experiment was carried out in pairs, each participant emailed one of the experiment supervisors and was allocated a specific time slot. In order to collect as many results as possible, in-house experiments took place to cater for participants who could not attend during the allocated time slots.

Participants at Nils Fredriksson Utbildning got information regarding the experiment through their student coordinator a few days prior to the experiment. When the supervisors had prepared the stations, they walked around and informed the students in the common areas. The students then showed up when they were able to participate in the experiment.

The number of female participants was 26 and the number of male participants was 22, as seen in Figure 4. The age range of participants was between 18-48 years of age. The distribution is shown in Figure 5. The most common background was engineering, see Figure 6. All backgrounds are presented in Section A.1 in Appendix A.

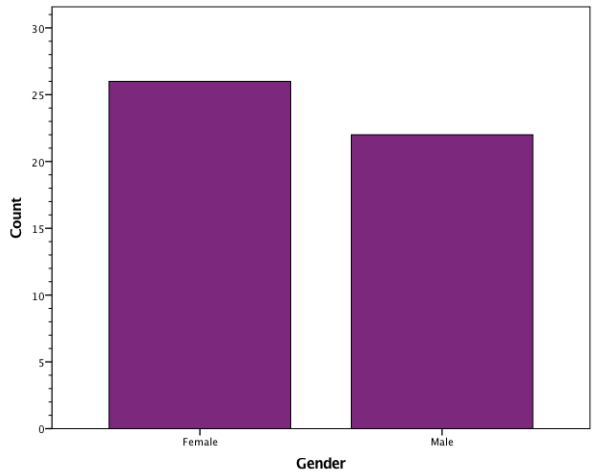



**Figure 4. The number of female and male participants. Slightly more females than males took part in the experiment.**

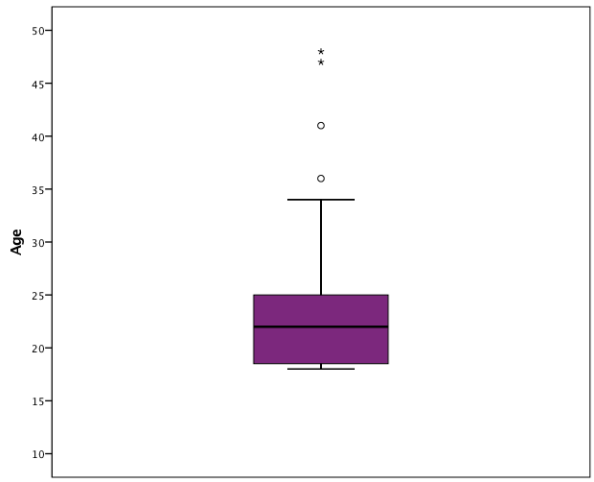



**Figure 5. The age of the participants. The median age is 22 years.**

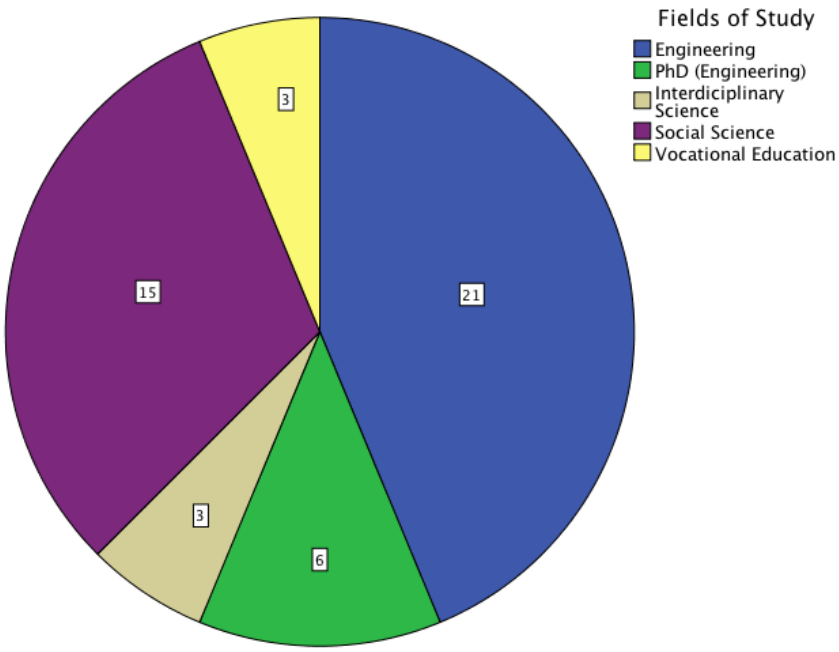



**Figure 6. The educational background of the participants. The majority of the participants have a background in engineering.**

Just below 60 % of the participants had been in contact with Mastermind before participating in the experiment, see Figure 7 for previous experience. In general, the participants had played during their childhood to some extent, however a few had played recently.

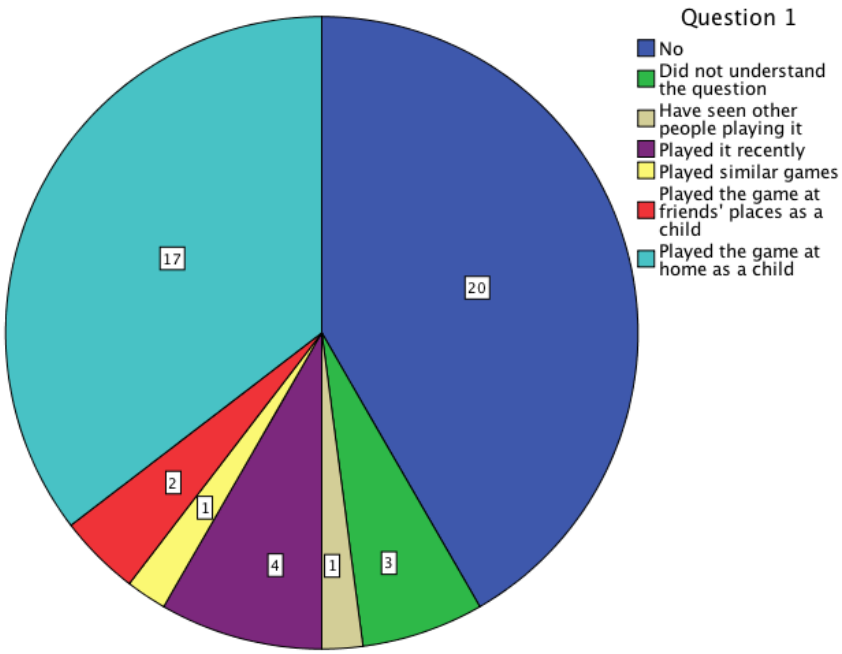



**Figure 7. The participants' previous experience of Mastermind. The majority had previous experience of the game.**

During the introduction the participants were briefed on capability assessments, how they are used and where they are applicable. In the questionnaire just below 70 % of the participants had never been in contact with capability assessments, see Figure 8. Among those who had previous experience of capability assessments the majority had performed capability assessments during their education, in their profession, in sports or in military service.

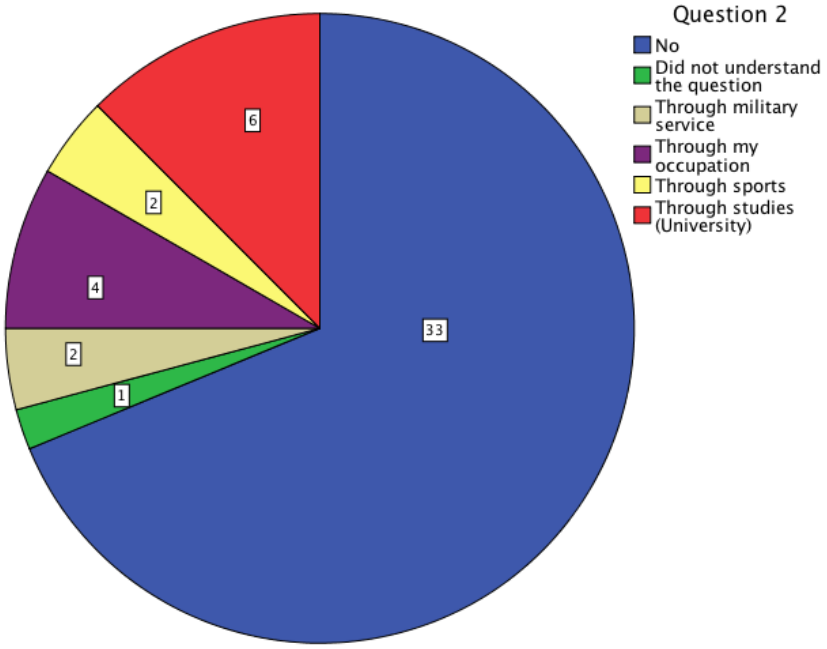



**Figure 8. The participants' previous experience of capability assessments. The majority had no previous experience of the game.**

## 5.3 Tasks

During the experiment the participants were asked to perform one of the two tasks at random, the tasks are described in Section 5.3. The two tasks represented Dependency I and Dependency II respectively. The dependency models are further described in Section 4.1.1.

The solution of each game board had been randomized prior the experiment. Each supervisor had six codes to alter between. All codes had four colours.

It was 14 pairs who did task I1/G1, 6 pairs who did task G2, and 4 pairs who did both.

### 5.3.1 I1/G1

In the I1/G1 task the participants had a game board and a code each. The results of the task were 1) the individual performance 2) the performance when the two individual performances were combined. This task has therefore two dependencies; 1) individual and therefore independent, 2) dependent on both actors. The design of the task means that the capability of one participant is not directly affecting the other participant's capability to solve their part of the task. Therefore, the participants are not directly dependent on each other to perform, but the result is dependent on both of them. See Figure 9.

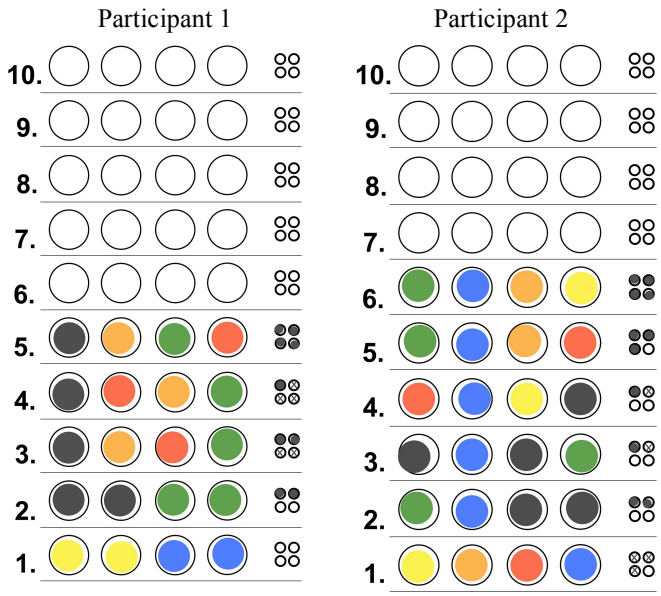



**Figure 9. An example result of the I1/G1 task. The results are 5 for Participant 1 and 6 for participant 2 for task I1. For task G1 the result is 11 (5+6).**

The implementation process of I1/G1 is as follows:

- The participants assess individually at which row they individually will solve their code.
- The participants assess individually at which row the pair will solve their codes.
- The participants assess in pairs at which row the pair will solve their codes

For each assessment, they are also asked to grade from 0 to 10 how:

- Certain they are to solve the code at the row from their assessment.
- Certain they are to solve the code +/- 1 row from their assessment.
- Certain they are to solve the code +/- 2 rows from their assessment.

The instructions to the participants can be found in Appendix B and the capability assessment form in Appendix C.

### 5.3.2 G2

In the G2 task the participants have a game board with a code each. After four rows the participants switch boards and consequently switch codes. During this task, the participants are dependent on each other in order to perform well. If one participant has used the first four rows illogical, it will be harder for the second participant to solve, hence this will affect the outcome. See Figure 10.

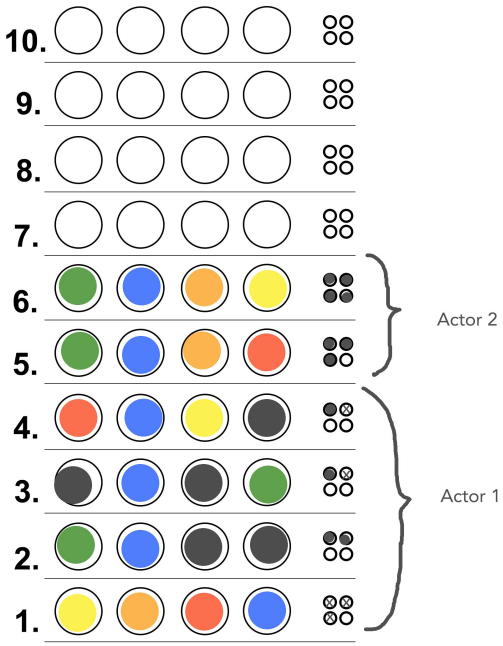



**Figure 10. A result for the task G2. The result is 6.**

The implementation process of G2 is as follows:

- The participants assess individually at which row the pair will solve their code.
- The participants assess in pairs at which row the pair will solve their code.

For each assessment, they are also asked to grade from 0 to 10 how:

- Certain they are to solve the code at the row from their assessment.
- Certain they are to solve the code +/- 1 row from their assessment.
- Certain they are to solve the code +/- 2 rows from their assessment.

The instructions to the participants can be found in Appendix B and the capability assessment form in Appendix C.

## 5.4 Conduction of experiment

In order to make all experiment sessions as similar as possible, the following checklist was observed:

1. All stations prepared.
2. Greeting the participants.
3. Short introduction to the thesis and to Mastermind.
4. Practice session for participants
5. Introduction to the task (I1/G1 or G2)
6. Participants perform the assigned task.
7. a) If the assigned task is completed when 20 minutes or more are remaining of the set time, next task is introduced and performed.
   b) If the assigned task is completed when less than 20 minutes are remaining, step 8 is followed.
8. Participants fill out the questionnaire.
9. Conclusion.

**All stations prepared**
During each session, one or two stations were used, depending on if two or four participants took part. One station set-up was as follows:

- Two sets of crayons; black, yellow, orange, red, blue and green.
- Two computers or tablets prepared with the version of Mastermind created by Lenssen (2015).
- Two pencils (participants)
- An erasable pencil (supervisors)
- For I1/G1: Five capability assessment forms, two game boards.
  For G2: Three capability assessments forms, two game boards.
- Two questionnaires in Swedish or English depending on the participants' preference
- A set of randomized codes

All forms, game boards and questionnaires were numbered before the participants arrived in order to follow one person's progress through the session.

**Greeting the participants**
The participants arrived to the room where the experiment was taking place. Before instructions were given, they were divided into pairs by the supervisors and assigned to a station.

**Short introduction to the thesis and Mastermind**
The participants got a short introduction to the thesis and Mastermind. The manuscript in Appendix B was used as a guideline for the introductions.

**Practice session for participants**
In order for the participants to fully understand Mastermind, they got 10 minutes to practice on a computer or tablet. During the 10 minutes' participants were allowed to ask the supervisors an unlimited amount of questions about the game. If the supervisors noticed that a participant did not understand the game, they explained again. The participants were given the advice to try several strategies to solve the game during the practice session.

**Introduction to the task (I1/G1 or G2)**
The participants were introduced to the task they were to perform, I1/G1 or G2, gradually. I1/G1 was carried out as follows:

1. The participants were asked to fill out the capability assessment form individually on how they thought they would perform individually.
2. The participants filled out the form, i.e. which row they assessed they would solve the code at and how confident they were in their assessment.
3. The participants were asked to fill out the capability assessment form individually on how they thought the pair would perform.
4. The participants individually filled out the form, i.e. which row they assessed they would solve the code at when combining the result of both participants' game boards and how certain they were on their assessment.
5. The participants were asked to fill out the capability assessment form together on how they thought the pair would perform.
6. The participants filled out the form in pairs, i.e. which row they assessed they would solve the code at when adding the result of both participants' game boards and how confident they were in their assessment.

G2 was carried out as follows:

1. The participants were asked to fill out the capability assessment form individually on how they thought the pair would solve the task.
2. The participants, individually, filled out the form, i.e. which row they assessed they would solve the code and how confident they were in their assessment.
3. The participants were asked to fill out the capability assessment form in pairs on how they thought the pair would solve the task.
4. The participants, in pairs, filled out the form, i.e. which row they assessed they would solve the code and how confident they were in their assessment.

**Participants perform the assigned task**
The participants performed the assigned task, either I1/G1 or G2. The tasks are described in Section 5.3.

**Participants fill out the questionnaire**
After the task was completed, the participants got a questionnaire in Swedish or English depending on their preference. If the participants had any queries, the supervisors explained the questions further. When the participants had filled out their questionnaires, the pair's capability assessments, game boards and questionnaires where put together.

**Conclusion**
The experiment ended when the participants had handed in their questionnaires and had received a movie ticket. Almost all participants decided to stay a few minutes to discuss what they had experienced. If the supervisors had queries regarding the participants' answers to the questionnaire or capability assessments, the participants were asked during this time.

# 6 Results

In this chapter all the results related to the conduction of the experiment are presented. All results from statistical tests are strictly related and confined to the framework of this thesis. The results are divided into three parts, depending on the source of information. Section 6.1 presents the results from the statistical testing, Section 6.2 presents results from the capability assessment form and Section 6.3 presents results from the questionnaire.

Detailed information about statistics is presented in Appendix A. Table 1 in Section A.2.2 presents the results from the statistical analysis related to the hypotheses, and Table 2 in Section A.2.2 presents descriptive statistics from which manual analysis can be made.

## 6.1 Hypothesis testing

**Hypothesis 1: There is no difference between capability assessments and performances.**
Five different comparisons were conducted through paired sample T-tests which all showed highly significant results, therefore hypothesis 1 is rejected. In other words, there is a statistical difference between capability assessments and performances. In addition, the calculations of Cohen's d showed exclusively large effect size in all comparisons.

**Hypothesis 2: There is no difference in accuracy between capability assessments made individually and in pairs for G1.**
One comparison was conducted through paired sample T-test which showed no significant results, therefore hypothesis 2 failed to be rejected. In other words, it is not possible to rule out the hypothesis. In addition, the calculations of Cohen's d showed a small effect size.

**Hypothesis 3: There is no difference in accuracy between capability assessments made individually and in pairs for G2.**
One comparison was conducted through paired sample T-test which showed no significant results, therefore hypothesis 3 failed to be rejected. In other words, it is not possible to rule out the hypothesis. In addition, the calculations of Cohen's d showed a small effect size.

**Hypothesis 4: There is no difference in accuracy for individual assessments between tasks (I1, G1, G2).**
A total of three comparisons were conducted (I1-G1, G1-G2, I1-G2), one of them through paired sample T-test (I1-G1) and two of them through independent sample T-tests (G1-G2, I1-G2). Two of the comparisons showed significant results (I1-G1, G1-G2), however this was not the case when comparing I1 and G2. Therefore, hypothesis 4 is rejected when comparing I1-G1 and G1-G2, but fails to be rejected when comparing I1 and G2. Calculations of Cohen's d showed a medium, large and small effect size respectively. Qualitative analysis of these results support the obtained quantitative results and will be further discussed in Section 7.1. In other words, there is a statistical difference in accuracy for individual assessments when comparing I1-G1 and G1-G2.

**Hypothesis 5: There is no difference in accuracy for pair assessments between tasks (G1, G2).**
One comparison was conducted through independent sample T-test which, after special consideration, showed no significant results, therefore hypothesis 5 failed to be rejected. In other words, it is not possible to rule out the hypothesis. These results will be discussed further in Section 7.1. since they showed a significant difference when not taking Levene's test into account and in addition the calculations of Cohen's d showed a large effect size.

**Hypothesis 6: There is no difference between capability assessments made individually and in pairs for G1.**
One comparison was conducted through paired sample T-test which showed no significant results, therefore hypothesis 6 failed to be rejected. In other words, it is not possible to rule out the hypothesis. In addition, the calculations of Cohen's d showed a small effect size.

**Hypothesis 7: There is no difference between capability assessments made individually and in pairs for G2.**
One comparison was conducted through paired sample T-test which showed no significant results, therefore hypothesis 7 failed to be rejected. In other words, it is not possible to rule out the hypothesis. In addition, the calculations of Cohen's d showed a small effect size.

**Hypothesis 8: There is no difference between performances of task I1 and G2.**
One comparison was conducted through independent sample T-test which showed no significant results, therefore hypothesis 8 failed to be rejected. In other words, it is not possible to rule out the hypothesis. In addition, the calculations of Cohen's d showed a small effect size.

## 6.2 Confidence in assessments

The responses to the Capability Assessments which did not follow common logic were removed for statistical purposes. This is mentioned further in Discussion, Section 7.

Table 2 presents results regarding the participants' confidence in assessments and should be interpreted as follows:

**Mean difference** is the difference between the mean value of capability assessments and the mean value of performance for each task. A positive result should be interpreted as an underestimation of actual capability, while a negative value should be interpreted as an overestimation of actual capability.

**Frequency, N (Percentage, %)** describes the number of participants who made a correct assessment, within the intervals of +/- 0, +/- 1, +/- 2 rows. It also describes the percentage of the sample for each interval who made a correct assessment.

**Confidence in assessment (Median)** describes the collective estimation about how confident the participants were in their assessments.

**Abbreviations** are as follows:

| | |
|---|---|
| ICAI1 | Individual Capability Assessment for I1 |
| ICAG1 | Individual Capability Assessment for G1 |
| ICAG2 | Individual Capability Assessment for G2 |
| GCAG1 | Pair Capability Assessment for G1 |
| GCAG2 | Pair Capability Assessment for G2 |

**Table 2. Comparison between the confidence of the participants and their performance.**

| Task | Mean difference (assessment-result) | Frequency, N (Percentage, %) | | | Confidence in assessment (Median) | | |
|---|---|---|---|---|---|---|---|
| | | +/- 0 | +/- 1 | +/-2 | +/- 0 | +/- 1 | +/-2 |
| **ICAI1** | 3.500 | 2 (5.55%) | 6 (16.66%) | 11 (30.55%) | 4 | 5.5 | 8 |
| **ICAG1** | 5.861 | 1 (2.77%) | 5 (13.88%) | 7 (19.44%) | 3 | 5 | 6.5 |
| **ICAG2** | 3.684 | 0 (0.00%) | 3 (15.00%) | 4 (20.00%) | 4 | 6 | 8 |
| **GCAG1** | 5.722 | 1 (5.55%) | 2 (11.11%) | 3 (16.66%) | 3 | 4 | 7 |
| **GCAG2** | 3.526 | 1 (5.88%) | 3 (17.64%) | 7 (41.17%) | 4 | 6 | 7.5 |

**Explanation of the results from Table 2:** Results for each participant's confidence can be interpreted as how many times out of ten tries will your assessment be correct and collectively this figure is comparable to the percentage of the sample who actually achieved a correct assessment. For example, ICAI1, the participants were confident that they would solve the code within an interval of +/- 0 rows, 40% of the time. i.e. will solve 4 out of 10 tries within the interval. However, they actually only reached a correct assessment at 5.55% of the time.

Table 2 shows exclusively overconfidence in the assessments. More detailed information about statistics can be found in Appendix A.

## 6.3 Questionnaire

Questions 3-7 are included in this section since they answer the research questions in part. Question 1 and 2 are not included in this section since they are presented in Section 5.2.

All free text answers have been divided into categories. The categories represent the intent of the replies. Where the answer has not been aligned with the question it has been put into the category "Did not understand the question". The participants had the opportunity to give free text answers to all questions, those who chose not to do so is put into the category "Chose not to respond".

**Question 3**

Did you experience it harder to do an individual assessment of a group task or to do a group assessment of a group task? Please put a cross on the line below where appropriate.

Individual assessment much harder. ——|——|——|——|——|——|——|——|——|—— Group assessment much harder.

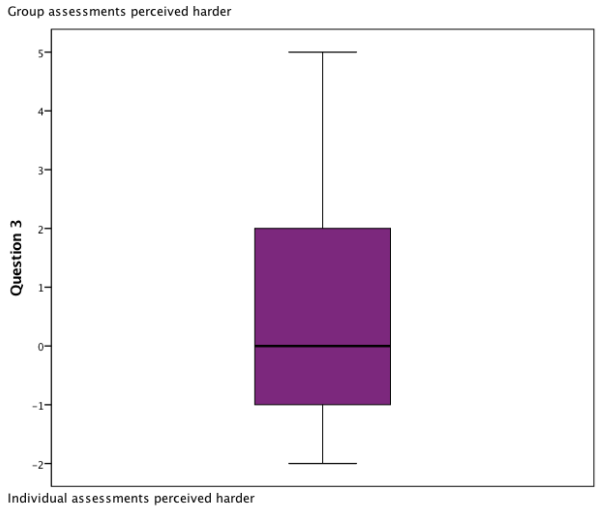



**Figure 11. Box plot of Question 3. The median shows that the sample found it equally difficult to assess the group task individually and in pairs. However, there is a larger dispersion among those who found it more difficult to assess in pairs.**

If a specific task was harder, please explain why:

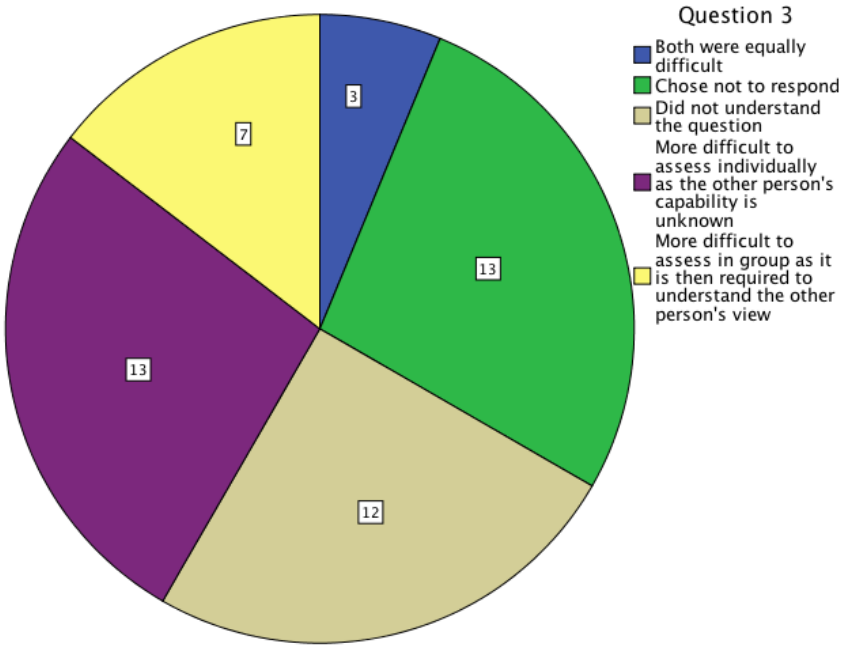



**Figure 12. Free text answers to Question 3 divided into categories. The majority of valid replies found it more difficult to assess the pair's capability individually as the other person's capability was unknown.**

**Question 4**

Did you think differently when reasoning in the individual assessments than in the group assessments? Yes No

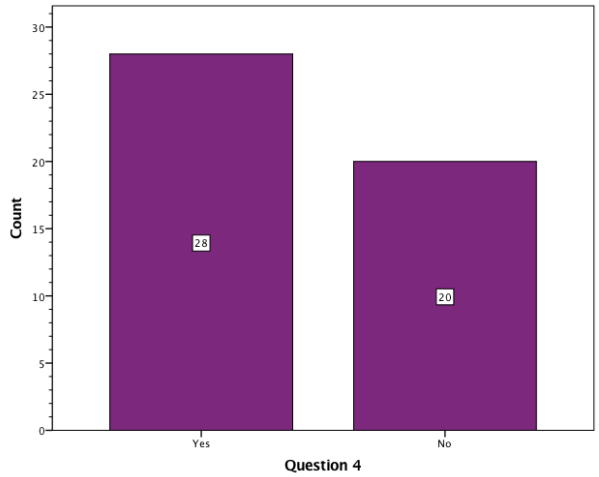



**Figure 13. Response to Question 4. The majority reasoned different between the assessments.**

If **Yes**, how did it differ? If **No**, how did you reason?

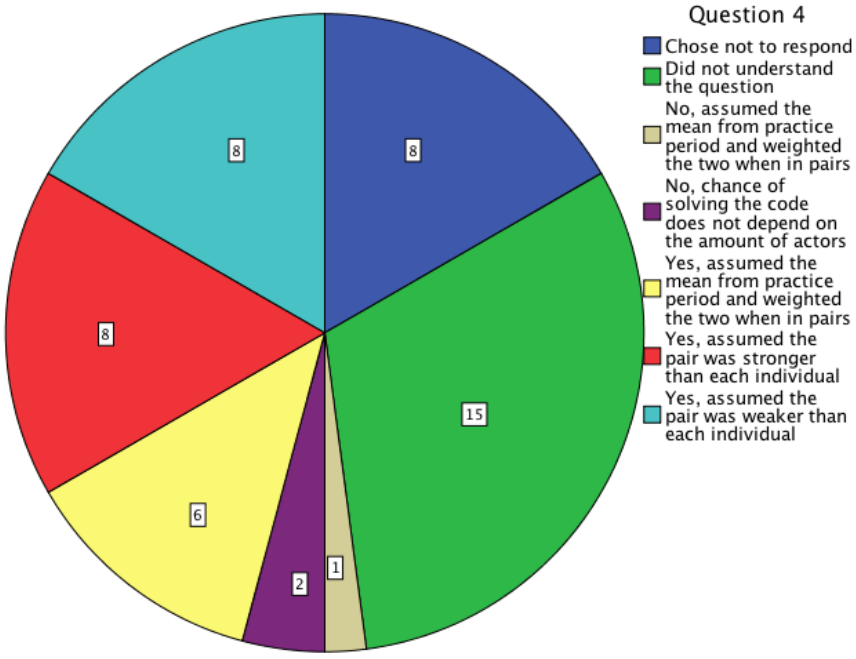



**Figure 14. Free text answers to Question 4 divided into categories. The majority of valid replies assumed that the pair was weaker than the individual.**

**Question 5**

Did you think differently in your individual assessments between the individual tasks and the group tasks? Yes No

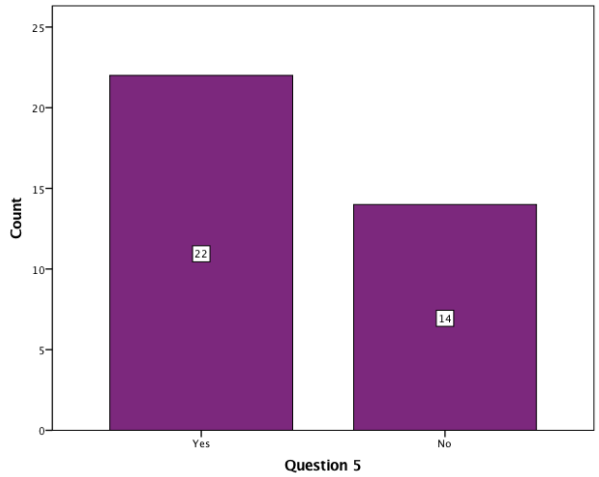



**Figure 15. Response to Question 5. The majority reasoned different between the assessments.**

If **Yes**, how did it differ? If **No**, how did you reason?

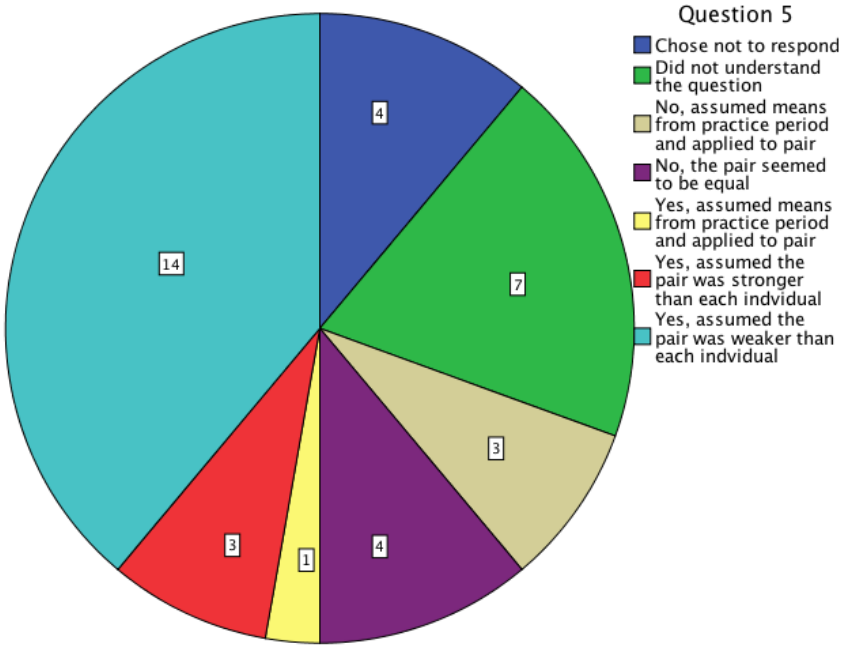



**Figure 16. Free text answers to Question 5 divided into categories. The majority of valid replies assumed the individual assessment of the group capability is less correct than the individual assessment of the individual capability.**

**Question 6**

Did you experience it harder to do the assessments of the first group task or the assessments of the second group task? Please put a cross on the line below where appropriate.

First group task (G1) much harder. ——————————— Second group task (G2) much harder.

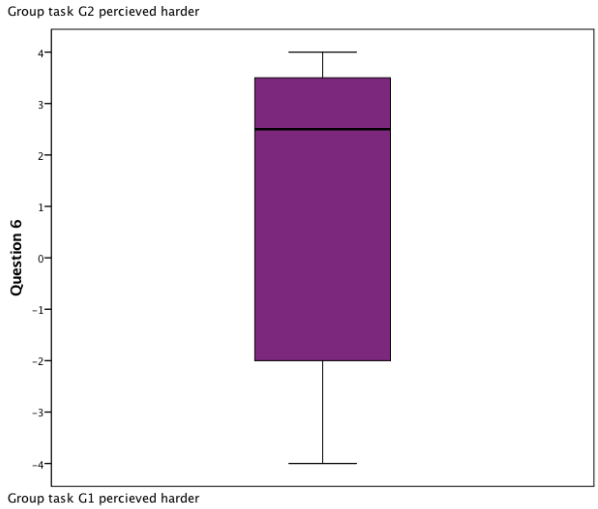



**Figure 17. Response to Question 6. The majority found it more difficult to assess task G2.**

If a specific task was harder, please explain why:

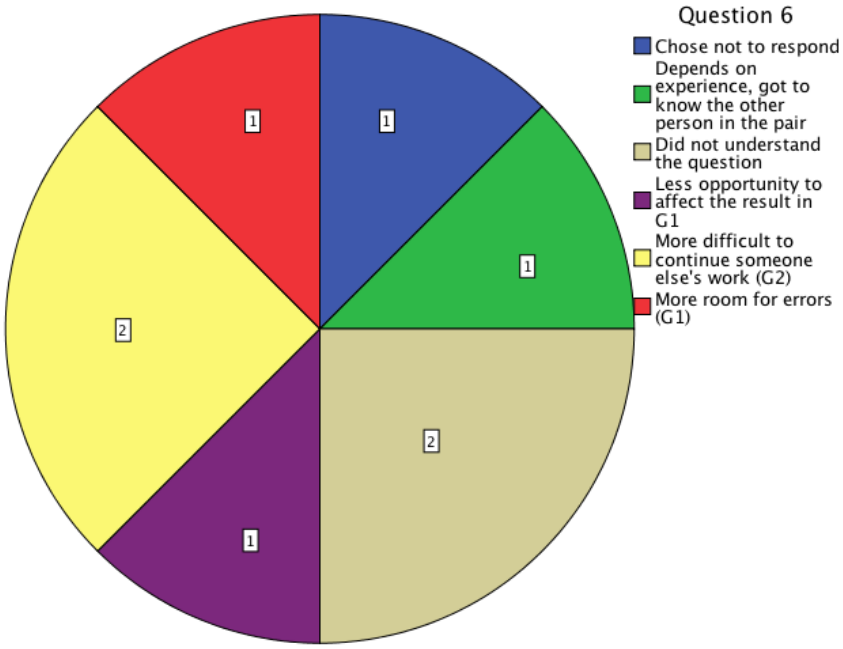



**Figure 18. Free text answers to Question 6 divided into categories. The majority of valid replies assumed it would be more difficult to continue their partner's work.**

**Question 7**
Other comments regarding the experiment (E.g level of difficulty of tasks, was it easy to follow the instructions and so on).

The intent in the majority of the replies to Question 7 was the participants were satisfied with the difficulty of tasks and found the experiment easy to follow. Two comments are considered relevant to present here:

- "To me it is important with a safety margin [when assessing capability, *authors' note*]. It feels safer to assess a level one know one is able to reach than to chance on a "good" performance when one might fail."

- "It is hard to know which strategy the other person has chosen. Working together with different strategies are difficult, but when the strategy of the other person is known it is easy [regarding task G2, *authors' note*]."

# 7 Discussion

In this Section the results, reliability, design issues and future implementation are discussed.

## 7.1 Results

The statistical testing showed that there was a significant difference between capability assessments and the actual performance for all tasks, with a systematic underestimation of the performance. This is not aligned with previous studies where the participants in general overestimate their capability (Mynttinen et al., 2009) (Dunning et al., 2004) (Fredriksson et al., 2011) (Vallone et al., 1990). In order to make a fair capability assessment, it is required to have relevant knowledge regarding the task being assessed, both regarding the task as it is and the actors performing it. The results suggest that the actors in the experiment did not acquire the relevant knowledge during the practice session and/or discussion with the partner. However, the participants of this study are overconfident in their assessments, which is similar to previous studies (Vallone et al., 1990) (Sung et al., 2010) (Fredriksson et al., 2011). It is the authors' of this thesis interpretation that underestimation of capability in previous studies has not been considered equal to overestimation. Instead, it has been considered better to make an underestimation than an overestimation, which is aligned with the participants' way of thinking in the experiment of this thesis.

One reason for the participants' systematic underestimation of their capability could be that although the participants were supposed to assess which row they thought they would solve the code *at,* they assessed which row they thought they would solve the code *by*. An indication of this way of thinking was given during the pilot studies when the participants explained how they were thinking during the experiment. In order to reduce the effects of this phenomenon, the supervisors specifically told the participants that not reaching the estimated row was considered equal to exceeding the estimated row. Despite of this encouragement, the participants continued to underestimate their capability.

In some cases the participants had estimated that they would solve the code by a fairly high number, row 11 or more. Despite of this, it was clear that they gave up to solve the code early on. With encouragement from their partner and/or supervisor, most of them solved the code with several rows to spare. This lack of performance confidence may also be a sign that it was prestige in solving the code before the assessed row.

Another reason could be that several participants wanted further explanation in addition to the instruction before filling in the capability assessment form. In the explanation, a number occurred in sentences similar to "if you assess you will solve the code at row 10, write 10". This, in addition to the knowledge that they could be given as many game boards as they required before solving the code, may have affected the participants to choose a higher row number than they otherwise would. This is what (Kahneman, 2011, p.92) describes as anchoring heuristics.

Where the participants rated how confident they were in their assessment, the individual assessments had a greater interval than the assessments made in pairs. One reason for this is that most pairs decided to compromise from the individual assessments and therefore used an averaged number. It was expressed in the questionnaires that the participants found it difficult to assess the pair's capability individually since the capability of the partner was unknown which could explain the wider spectra of the individual assessments. However, several

participants noted that they found it more difficult to assess in pairs since it required understanding of how the other participant rated their individual and the pair's skill.

When assessing which row the pair would solve the code at, the pair often reasoned that because they were two, the results could either be better or worse than the individual results. They rarely mentioned that the performance could be the same, i.e. one participant could make a bad performance, and the second a good performance weighing up for the first. This was expressed in the discussions between the participants during the experiment. It was also expressed in the questionnaire that the participants considered the pair's capability being weaker than the individual capability, both when assessing individually and in pairs. This is aligned with the free text answers from the questionnaire where the majority of the relevant replies considered it being more difficult to assess the group tasks than the individual task. It also seems like the participants felt more 'secure' by underestimating their capability. A comment from the questionnaire confirms this by stating that it was important to the participant to have a safety margin.

A qualitative analysis was made of hypotheses 4 and 5 where the accuracy of individual and group capability assessments was tested. The qualitative analysis of the individual assessments showed that the tasks are ranked I1, G2 and G1 in accuracy. This is based on that there are significant differences between I1 and G1 as well as between G2 and G1 and that the accuracy mean is lower for I1 than for G2. Hypothesis 5 was not rejected as Levene's test showed that there was not a significant difference between G1 and G2. However, a qualitative analysis shows that in capability assessments made in pairs G2 tend to be more accurate than G1.

Before the experiments, it was the authors' belief that the pair assessments would be more accurate than the individual assessments for pair tasks. However, it was not possible to reject the hypotheses that there were no differences in accuracy between the capability assessments made individually or in pairs for G1 or G2. This suggests that despite more knowledge about the pair's common capability, it does not improve the validity of the assessments. Klein and Canderwood (1991) suggest that experts react to one or several cues when making decisions. It is possible that the participants knew too little about their capability in order to make an accurate assessment, hence no cues where given to be able to foresee potential solutions. The activity Mastermind might also be of a nature where no cues are given as it lacks complexity. Instead, the participants probably used the decision tree model, which evaluates proficiencies and constraints for all options before decisions are made (Klein & Calderwood, 1991). Using the decision tree model take more time than reacting to cues, which could be a reason why the participants spent a disproportionate amount of time at the first few rows noted in the pilot studies.

The main argument to include task I1 was to investigate if there was any difference between individual performance and group performance when it is required to understand how another person has approached a problem. The results show that there might be no difference between the two. The authors found this surprising because it was anticipated that G2 would take longer to perform since it required the second actor to understand the strategy used by the first actor. Although time was not measured, the authors found that there was hardly any time difference between G2 and I1.

## 7.2 Reliability

The experimental design is thoroughly described in this thesis. It is the authors' belief that the experiment is possible to recreate with similar results. However, due to the nature of Mastermind and the design of the two tasks (G1 and G2), it may be difficult to replicate the experiment with other activities than Mastermind. The way G1 is designed the performance is often twice the number of rows than for G2. As G1 and G2 are different, it is not possible to draw any conclusions how the performances of the two dependency models differ, other than their relation to the relevant capability assessment.

## 7.3 Design issues

Mastermind was considered a suitable activity for this experiment since there were many advantages. For example, tasks for the two different dependencies presented in Section 4.1.1 could be designed. However, the very nature of Mastermind made the two multi-actor dependency tasks vary in amount of rows. It is apparent that G1, where the participants combined their results would generate a larger amount of rows than G2. For example, the mean of the capability assessments for G1 is 20 and the mean of performances for G1 is 14, while the means are 10.6 and 6.9 for the capability assessments and performances for G2 respectively. This means that the performances of G1 and G2 were not as comparable as initially planned. Instead the performance and capability assessments are only comparable if accuracy is introduced. Despite of this design issue, the result from I1 and G2 is comparable as they have similar results. However, if this is due to the dependencies or the design is difficult to distinguish.

The authors did not register the results from the practice sessions, but it was apparent that it was more challenging for the participants than in the practical experiment. One of the obvious reasons for this is that it was often their first encounter with Mastermind. Another difference between the practice session and the practical experiment is that the participants used an Internet based Mastermind where all colour combinations were allowed during the practice session. That means that a code could be four reds, or three blues and one green, for example. However, in the practical experiment it was always four different colours. The participants were unaware of this difference and this may have affected the participants' perception of difficulty level. Another reason for the underestimated performances by the majority of the participants is the difficulty of assessing when biased by previous results. Kahneman and Klein (2009) describe that it is common to underestimate one's performance if there is a history of failure, while it is common to overestimate one's performance if there is a history of success.

Challenges with the questionnaire were found during the pilot study, consequently the supervisors offered to explain the questions for each participant. Despite of this, it was apparent in some free text answers that the participant had not understood the question. To some extent this was more common among the younger participants. Despite several reconstructions to make the questionnaire easier to follow, it was a challenge for the participants to understand. In addition to typos in Question 3b and 4, the authors believe the questionnaire was hard to understand because the questions were similar and easily misinterpreted. For example, it was apparent in the free text replies that individual capability assessment and individual performance were often interpreted as interchangeable. The aim of Question 4 and 5 was to investigate if the participants made a difference in their assessments depending on if the assessments were made individually or in pairs and if the task were performed individually or in group respectively. It was found difficult to develop these

questions without giving examples that probably had affected the replies. Instead several responses were considered irrelevant.

The capability assessment form was considered to suit its purpose. However, two interesting discoveries were made. Firstly, a few of the participants chose 0 on the 0-10 scale on how confident they were that they would solve the code on the assessed row. As it was discovered after the participants had left the premises no follow up was made. It is therefore the authors' interpretation that the participants did not change row since they were equally unconfident they would sole the code at any other row. Secondly, a few participants noted that they were less confident when the interval of rows was larger. The participants explained this as they considered it more likely for them to sole the code at that row or not at all in the vicinity. This does not follow common logic and such results were not included in the analysis.

An unexpected issue was that a few participants were colour blind. When presented with the suggestion of a code made of patterns instead, the participants chose to continue with colour codes. They mentioned that they believed that they had no issue to solve the code as they only had to keep track on which shade of colour that is supposed to be where rather than identifying which colour each dot was. It is the authors' interpretation that the performance of these participants does not differ from other participants' results.

## 7.4 General discussion

Overestimating or underestimating the capability where multi-actor dependencies are present may cause consequences for either dependent actors or the result. For example, an underestimation of a specific actor's capability may cause misplaced resources.

In the making of an assessment of capability, there are always assumptions being made regarding the conditions, circumstances and dependencies. It is difficult, but essential, to pinpoint which these are as without traceable assumptions it is possible to assume all other actors remain unaffected in the scenario of the capability assessment. Further, where many actors are involved it is challenging to pinpoint who is responsible for the collective capability.

During the experiment it was observed that pairs already acquainted with each other had internal power structures that affected their joint capability assessments. Pairs that were unknown to each other were observed more polite and reached a more general consensus. However, when one of the two participants unknown to each other was more outgoing, the capability assessment tended to be more aligned to this participant's individual assessment.

It is crucial to have relevant knowledge about the task being assessed, both regarding the task as it is and the actors performing it in order to make a fair capability assessment. Especially as one actor rarely are able to have more than a fragmented overview of the situation.

## 7.5 Future implementation of capability assessments

The world grow more and more complex with an increasing amount of dependencies between actors. The increasing amount of actors and the dependencies in between, increase the risk of misinterpretation of responsibilities and capabilities. Therefore, it will be crucial in the long run to conduct capability assessments across boundaries. This is not only relevant to risk and vulnerability analyses within public agencies but also at the intersection between public and

private responsibilities.

In reality, the challenges faced by actors doing capability assessments with multi-actor dependencies are more complex than the experiment conducted in this thesis. An overestimation of a joint capability may ultimately result in loss of lives. While an underestimation may not result in loss of lives directly, it may result in misplaced resources which indirect may result in loss of lives. Therefore, it is critical to assess capability without overestimation or underestimation.

Capability assessments in the form used today is a fairly new phenomenon, and it will take a while until the practical methods are implemented. However, in a not to distant future, capability assessments across established boundaries will be required to ensure resilience towards unknown risks.

The results of this study suggest that although people find it more difficult to assess capability for Dependency II tasks, i.e. where actors are dependent on each other to perform, than for Dependency I tasks, i.e. where the actors are not dependent on each other to perform, assessments for Dependency II tasks tend to be more accurate. A reason for this could be that as it is perceived harder, actors are more thorough when assessing capability for Dependency II. Also, it might be more apparent that the actors are striving towards the same goal. With a more substantial common goal and that one actor's performance affect all other actors, the actors might perceive their contribution as more important. Which result in a more thorough performance by the relevant actors. This suggests that tasks with multi-actor dependencies are vulnerable if they are performed with little or no communication.

Although communication is not the focus of this study, it is apparent that it is crucial for future capability assessments with multi-actor dependencies that requires more than a fragmental picture regarding the other actor's responsibility and capability.

# 8 Conclusion

In this Section the conclusions from the research questions are provided. The aim of the secondary research questions has been to provide a nuanced conclusion to the overall research question.

In this Section, the performed tasks are noted as I1, G1 and G2 respectively. G1 represent Dependency I and G2 represent Dependency II. The tasks are further described in Section 5.3.

## 8.1 Secondary research questions

*Do the capability assessments match the actual performance?*

No, the capability assessments do not match the actual performance since there was a significant difference between the two for all tasks. The participants, both individually, and when assessing in pairs, underestimate their capability for all tasks.

*Is there a difference in accuracy between capability assessments performed individually and in pairs?*

The statistical testing does not show a significant difference in accuracy between capability assessments performed individually and in pairs for either task. The participants tended to be more confident in their pair assessments for Dependency II and in their individual assessment for Dependency I.

*Is there a difference between capability assessments depending on if the assessment was made individually or in pairs?*

The statistical testing does not show a significant difference between the capability assessments depending on if they were made individually or in pairs. However, the individual assessments tend to be more dispersed than the pair assessments.

*Is there a difference between how well the capability assessments match the actual performance depending on if the task was performed individually or in pairs?*

One of the two relevant parts of the hypothesis is rejected. There is a significant difference in accuracy between I1 and G1 when the capability assessments are made individually, where the individual capability assessments for I1 are more accurate than for G1. However, there is not a significant difference between individual capability assessments for I1 and G2.

*Is there a difference between how well the capability assessments match the actual performance depending on how the multi-actor dependencies were designed?*

There is a statistical difference in accuracy between G2 and G1 when the capability assessments were made individually, showing that the assessments for G2 were more accurate than for G1. There is not a statistical difference in accuracy between G2 and G1 when the capability assessments were made in pairs. However, capability assessments for G2 tended to be more accurate than for G1.

*Do performances where multi-actor dependencies are present, differ from individual performances?*

There is not a significant difference between the performance of I1 and G2. Since the design of G1 differ from I1 and G2 it is not possible to include G1 in this comparison.

## 8.2 Overall research question

*How does multi-actor dependencies affect capability assessments?*

This study shows that multi-actor dependencies affect:

- *The participants' sense of confidence.* Pair assessments tend to be more confident than individual assessments for Dependency I and tend to be less confident for Dependency II. There are no significant differences in accuracy between assessments made in pairs or individually.

- *The accuracy between individual assessments and tasks.* A qualitative analysis suggests that it is easier for an individual to assess the individual task and the Dependency II task than the Dependency I task.

- *The accuracy between pair assessments and tasks.* A qualitative analysis shows a tendency that the Dependency II task is more accurate in pair assessments than the Dependency I task. Although this need to be confirmed by future studies.

- *The participants' sense of security.* Although the participants were instructed that underestimation and overestimation of capability were treated equally, a majority of the participants underestimated their capability to avoid perceived failure. This phenomenon exists in all capability assessments but implies greater consequences when multiple actors are involved.

# 9 Future research

LUCRAM and PRIVAD are already conducting research with the aim to improve capability assessments in terms of comparability, user friendliness and method development. It is of special interest to involve multi-actor dependencies in their future research.

The focus of this thesis is how multi-actor dependencies affect capability assessments and what this may imply for risk and vulnerability analyses developed with several actors. There are limitations to the applicability of the results in this thesis, mainly because of the simplifications made in order to draw any conclusions. To be able to confirm that the findings of this thesis is applicable in more complex settings, the following topics are proposed to be researched further:

**Evaluate to which extent an overestimation or underestimation affect resource spending and the consequences that follow when multi-actor dependencies are present.**
In this thesis it has been established that overestimating or underestimating capability may affect the coalition between actors, but not to which extent. A future study could bring clarity in how overestimation and underestimation of one or two actors' capability affect the joint capability.

**Evaluate how the level of complexity in tasks affects accuracy of capability assessments.**
The tasks performed by the participants in this thesis were simplified in order to be able to draw conclusions. However, there are limitations on how to apply the results in settings outside of the experiment. In order to be able to draw further conclusions, more complex tasks are required to be explored.

**Evaluate how power structures affect multi-actor dependencies.**
During this thesis a tendency was found that internal power structures affect mutual achievements where multi-actor dependencies are relevant. A future study could focus on how this is expressed and how to avoid the internal power structures to affect multi-actor capability assessments.

**Evaluate how experience affects multi-actor dependencies**
It was considered outside the scope of this thesis to discuss if/how experience affect capability assessments where multi-actor dependencies are present. However, previous studies have shown a tendency that inexperienced and experienced participants perceive tasks differently. This may impact capability assessments where multi-actor dependencies are present.

# 10 References


Aven, T., 2011. On Some Recent Definitions and Analysis Frameworks for Risk, Vulnerability, and Resilience. *Risk Analysis*, pp.515-22.

Balcetis, E., Dunning, D. & Miller, R.L., 2008. Do Collectivists Know Themsselves Better Than Individualists? Cross-Cultural Studies of the Holier Than Thou Phenomenon. *Journal of Personality and Social Psychology*, pp.1252-67.

Bhatta, G., 2003. Intent, risks and capability: some considerations on rethinking organizational capability. *International Review of Administrative Sciences* , pp.401-18.

Cunningham, D.W. & Wallraven, C., 2012. *Experimental Design: from user studies to psychophysics*. Boca Raton: CRC Press.

Dekker, S. & Hollnagel, E., 2004. Human factors and folk models. *Cogn Tech Work*, pp.79-86.

Dunning, D., Heath, C. & Suls, J.M., 2004. Flawed Self-Assessment: implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest*, pp.69-106.

Fredriksson, U., Villalba, E. & Taube, K., 2011. Do Students Correctly Estimate Their Reading Ability? A Study of Stockholm Students in Grades 3 and 8. *Reading Psychology*, pp.301-21.

Höst, M., Regnell, B. & Runeson, P., 2011. *Att genomföra examensarbete*. 1st ed. Lund: Studentlitteratur.

Haimes, Y.Y., 2006. On the Definition of Vulnerabilities in Measuring Risks to Infrastructures. *Risk Analysis*, pp.293-96.

Halkjelsvik, T., Jørgensen, M. & Teigen, K.H., 2011. To Read Two Pages, I Need 5 Minutes, but Give Me 5 Minutes and I will Read Four: How to Change Productivity Estimates by Inverting the Question. *Applied Coginitive Psychology*, pp.314-23.

Heylighen, F., Cilliers, P. & Gershenson, C., 2007. Complexity and Philosophy. *Complexity, Science and Society*.

Hills, A., 2005. Insidious Environments: Creeping Dependencies and Urban Vulnerabilities. *Journal of Contingencies and Crisis Management*, March. pp.12-20.

Körner, S. & Wahlgren, L., 2006. *Statistisk dataanalys*. 4th ed. Lund: Studentlitteratur.

Kahneman, D., 2011. *Thinking Fast and Slow*. 1st ed. New York: Farrar, Straus and Giroux.

Kahneman, D. & Frederick, S., 2001. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin & D. Kahneman, eds. *Heuristics and Biases - The Psychology of Intuitive Judgement*. 1st ed. Cambridge University Press. pp.49-81.

Kahneman, D. & Klein, G., 2009. Conditions for Intuitive Expertise: A Failure to Disagree. *American Psychologist* , September. pp.515-26.

Kahneman, D. & Tversky, A., 1973. On the psychology of prediction. *Psychological Review*, July. pp.237-51.

Kahn, K.S., Kunz, R., Kleijnen, J. & Antes, G., 2003. Five Steps to Conducting a Systematic Review. *Journal of the Royal Society of Medicine*, March. pp.118-21.

Klein, G., 2008. Naturalistic Decision Making. *Human Factors: The Journal of the Human Factors and Ergonomics Society* , June. pp.456-60.

Klein, G.A. & Calderwood, R., 1991. Decision Models: Some Lessons From the Field. *IEEE Transactions on Systems, Man, and Cybernetics*, September/October. pp.1018-26.

Lenssen, P., 2015. *Games for the brain*. [Online] Available at: http://www.gamesforthebrain.com/game/guesscolors/ [Accessed 14 September 2015].

Lindbom, H., Tehler, H. & Eriksson, K., 2015a. The capability concept – On how to define and describe capability in relation to risk, vulnerability and resilience. *Reliability Engineering and System Safety*, pp.42-54.

Lindbom, H., Tehler, H., Frykmer, T. & Uhr, C., 2015b. How can the usefulness of capability assessments be improved? Lund, 2015b. Lund University.

MSB, 2014. *Konsekvensutredning för föreskrift om kommuners och landstings risk- och sårbarhetsanalyser*. Konsekvensbeskrivning. Myndigheten för samhällsskydd och beredskap.

MSBFS 2010:6. *Myndigheten för samhällsskydd och beredskaps föreskrifter om kommuners risk- och sårbarhetsanalyser.*

MSBFS 2015:4. *Myndigheten för samhällsskydd och beredskaps föreskrifter om kommuners risk- och sårbarhetsanalyser.*

MSBFS 2015:5. *Myndigheten för samhällsskydd och beredskaps föreskrifter om kommuners risk- och sårbarhetsanalyser.*

Mynttinen, S. et al., 2009. Self-assessed driver competence among novice drivers – a comparison of driving test candidate assessments and examiner assessments in a Dutch and Finnish sample. *Journal of Safety Research*, pp.301-09.

Palmqvist, H. et al., 2012. *Utveckling av förmågebedömningar*. Lund: LUCRAM Lund University Centre for Risk Assessment and Management.

Palmqvist, H., Tehler, H. & Shoaib, W., 2014. How is capability assessment related to risk assessment? Evaluating existing research and current application from a design science perspective. In *Probabilistic Safety Assessment and Management conference*. Lund, 2014. Probabilistic Safety Assessment and Management.

SFS 2006:544. *Lag om kommuners och landstings åtgärder inför och vid extraordinära händelser i fredstid och höjd beredskap.*

Sundvik, L. & Lindeman, M., 1998. Acquaintanceship and the Discrepancy between Supervisor and Self-Assessments. *Journal of Social Behavior and Personality*, pp.117-26.

Sung, Y.-T., Chang, K.-E., Chang, T.-H. & Yu, W.-C., 2010. How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolecence*, pp.135-45.

Vallone, R.P., Griffin, D.W., Lin, S. & Ross, L., 1990. Overconfident Prediction of Future Actions and Outcomes by Self and Others. *Journal of Personality and Social Psychology* , pp.582-92.

# Appendix A - Statistics

*In this Appendix all statistics are collected.*

## A.1 Participants

In this Section the background information regarding the participants is presented.

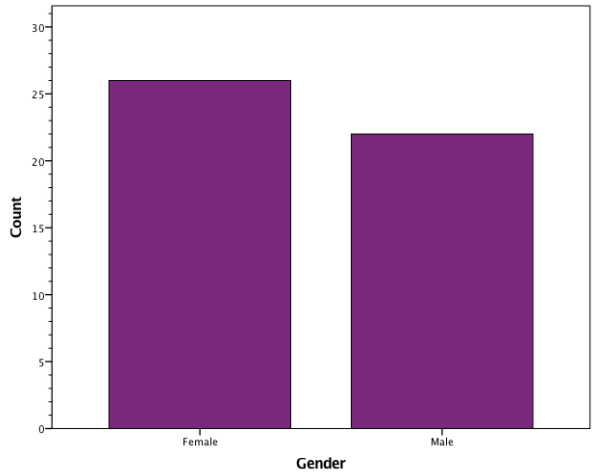



**Figure 19. Gender distribution.**

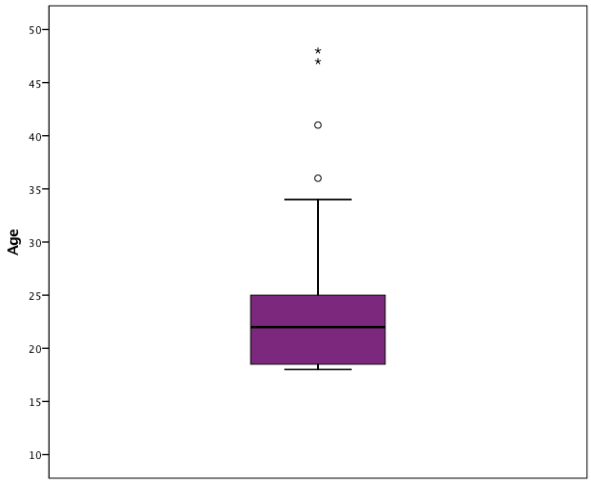



**Figure 20. Age distribution.**

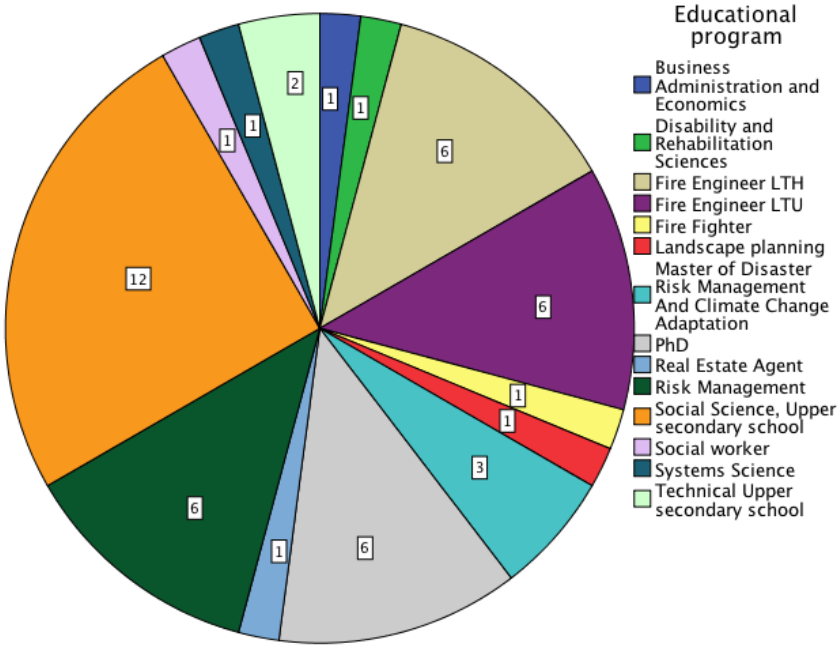



**Figure 21. Educational program.**

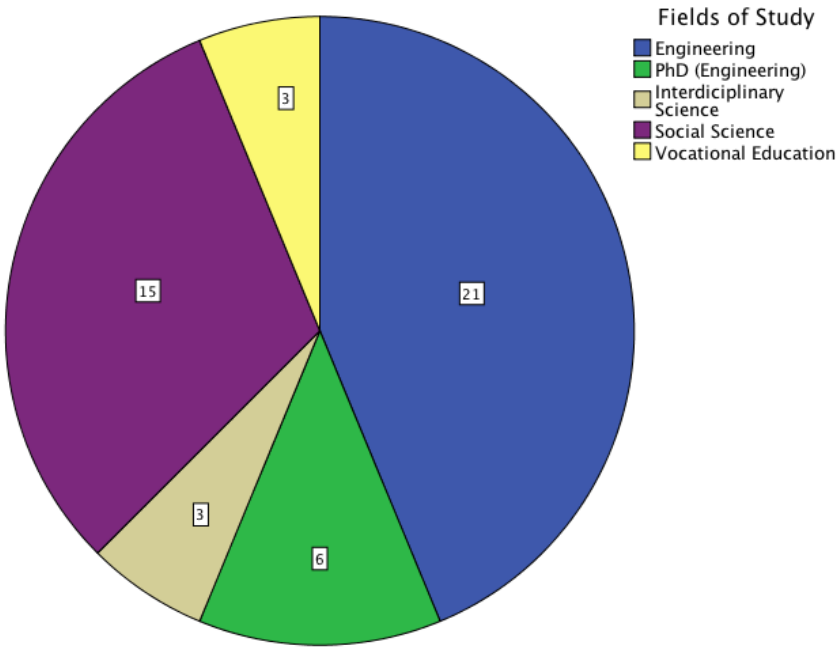



**Figure 22. Represented fields of study.**

## A.2 Compilation of results from statistical analysis

In this Section are the results from the statistical testing presented.

### A.2.1 Abbreviations

During the data collection and the statistical analyses, it was necessary to adopt abbreviations for specific data points. The following naming system was used:

WXYZ, where:

W = D, difference between capability assessment and performance. Excluded where CA or P is used.
X = I or G, describes if the task or assessment was made individually (I) or in pairs (G).
Y = CA or P, which describes if it is a capability assessment (CA) or the actual performance (P). Excluded where D is used.
Z= I1, G1 or G2, which describes which task it concerns, the individual task (I1), the type 1 dependency (G1), or the type 2 dependency (G2).

The naming system The abbreviations used are as follows:

| | |
|---|---|
| ICAI1 | Individual Capability Assessment for I1 |
| ICAG1 | Individual Capability Assessment for G1 |
| ICAG2 | Individual Capability Assessment for G2 |
| GCAG1 | Pair Capability Assessment for G1 |
| GCAG2 | Pair Capability Assessment for G2 |
| DII1 | Difference between Individual Capability Assessment and Performance for I1 |
| DIG1 | Difference between Individual Capability Assessment and Performance for G1 |
| DIG2 | Difference between Individual Capability Assessment and Performance for G2 |
| DGG1 | Difference between Pair Capability Assessment and Performance for G2 |
| DGG2 | Difference between Pair Capability Assessment and Performance for G2 |
| PI1 | Performance for I1 |
| PG1 | Performance for G1 |
| PG2 | Performance for G2 |

### A.2.2 Hypothesis testing

Table 3 presents the results from the statistical analysis related to the hypotheses.

**Table 3. Compilation of the statistical analysis of the hypotheses.**

| **Hypothesis** | **Statistical test** | **Significance level** | **$H_0$ Rejected/Not rejected** | **Cohen's d** |
|---|---|---|---|---|
| **Hypothesis 1** | **There is no difference between capability assessments and performances** | | | |
| **ICAI1 - PI1** | Paired sample T-test | .000 | Rejected | 0.7991 |
| **ICAG1 - PG1** | Paired sample T-test | .000 | Rejected | 0.7837 |
| **ICAG2 - PG2** | Paired sample T- | .000 | Rejected | 1.6893 |

| | | | | |
|---|---|---|---|---|
| | test | | | |
| **GCAG1 - PG1** | Paired sample T-test | .002 | Rejected | 0.9009 |
| **GCAG2 - PG2** | Paired sample T-test | .000 | Rejected | 1.6465 |
| **Hypothesis 2** | **There is no difference in accuracy between capability assessments made individually and in pairs for G1** | | | |
| **DIG1 - DGG1** | Paired sample T-test | .776 | Not rejected | -0.05 |
| **Hypothesis 3** | **There is no difference in accuracy between capability assessments made individually and in pairs for G2** | | | |
| **DIG2 - DGG2** | Paired sample T-test | .268 | Not rejected | -0.2512 |
| **Hypothesis 4** | **There is no difference in accuracy for individual assessments between tasks (I1, G1, G2)** | | | |
| **DII1 - DIG1** | Paired sample T-test | .005 | Rejected | 0.5943 |
| **DIG1 - DIG2** | Independent sample T-test | .031 (.016 Levene's test adaption) | Rejected | -0.7766 |
| **DII1 - DIG2** | Independent sample T-test | .707 (.658 Levene's test adaption) | Not rejected | 0.1188 |
| **Hypothesis 5** | **There is no difference in accuracy for pair assessments between tasks (G1, G2)** | | | |
| **DGG1 - DGG2** | Independent sample T-test | .027 (.059 Levene's test adaption) | Special consideration, Not rejected | -0.8519 |
| **Hypothesis 6** | **There is no difference between capability assessments made individually and in pairs for G1** | | | |
| **ICAG1 - GCAG1** | Paired sample T-test | .776 | Not rejected | 0.0504 |
| **Hypothesis 7** | **There is no difference between capability assessments made individually and in pairs for G2** | | | |
| **ICAG2 - GCAG2** | Paired sample T-test | .494 | Not rejected | 0.1558 |
| **Hypothesis 8** | **There is no difference between performances of task I1 and G2** | | | |
| **PI1 - PG2** | Independent sample T-test | .785 (.767 Levene's test adaption) | Not rejected | 0.0880 |

Table 4 presents descriptive statistics from which manual analysis can be made.

**Table 4. Descriptive statistics.**

| Hypothesis | Statistical test | Mean difference | Mean variable 1 | Mean variable 2 | N variable 1 | N variable 2 |
|---|---|---|---|---|---|---|
| **Hypothesis 1** | **There is no difference between capability assessments and performances** | | | | | |
| **ICAI1 - PI1** | Paired sample T-test | 3.500 | 10.611 | 7.111 | 36 | 36 |
| **ICAG1 - PG1** | Paired sample T-test | 5.861 | 20.083 | 14.222 | 36 | 36 |
| **ICAG2 - PG2** | Paired sample T-test | 3.684 | 10.737 | 7.053 | 19 | 19 |
| **GCAG1 - PG1** | Paired sample T-test | 5.722 | 19.944 | 14.222 | 18 | 18 |
| **GCAG2 - PG2** | Paired sample T-test | 3.526 | 10.579 | 7.053 | 19 | 19 |
| **Hypothesis 2** | **There is no difference in accuracy between capability assessments made individually and in pairs for G1** | | | | | |
| **DIG1 - DGG1** | Paired sample T-test | -.139 | -5.86 | -5.72 | 36 | 36 |
| **Hypothesis 3** | **There is no difference in accuracy between capability assessments made individually and in pairs for G2** | | | | | |
| **DIG2 - DGG2** | Paired sample T-test | -.158 | -3.68 | -3.53 | 19 | 19 |
| **Hypothesis 4** | **There is no difference in accuracy for individual assessments between tasks (I1, G1, G2)** | | | | | |
| **DII1 - DIG1** | Paired sample T-test | 2.361 | -3.50 | -5.86 | 36 | 36 |
| **DIG1 - DIG2** | Independent sample T-test | -3.636 | -7.54 | -3.90 | 28 | 20 |
| **DII1 - DIG2** | Independent sample T-test | .400 | -3.50 | -3.90 | 36 | 20 |
| **Hypothesis 5** | **There is no difference in accuracy for pair assessments between tasks (G1, G2)** | | | | | |
| **DGG1 - DGG2** | Independent sample T-test | -3.688 | -7.21 | -3.53 | 14 | 19 |
| **Hypothesis 6** | **There is no difference between capability assessments made individually** | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | **and in pairs for G1** | | | | | |
| **ICAG1 - GCAG1** | Paired sample T-test | .1389 | 20.083 | 19.944 | 36 | 36 |
| **Hypothesis 7** | **There is no difference between capability assessments made individually and in pairs for G2** | | | | | |
| **ICAG2 - GCAG2** | Paired sample T-test | .1000 | 10.600 | 10.500 | 20 | 20 |
| **Hypothesis 8** | **There is no difference between performances of task I1 and G2** | | | | | |
| **PI1 - PG2** | Independent sample T-test | .2066 | 7.259 | 7.053 | 27 | 19 |

## A.3 Raw statistical data

In this Section are the results from the statistical analyses made in SPSS presented.

### A.3.1 Hypothesis 1

**Table 5. SPSS results for Hypothesis 1.**

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | ICAI1 | 10.611 | 36 | 3.1738 | .5290 |
| | PI1 | 7.111 | 36 | 2.7021 | .4504 |
| Pair 2 | ICAG1 | 20.083 | 36 | 6.2307 | 1.0384 |
| | PG1 | 14.222 | 36 | 3.4235 | .5706 |
| Pair 3 | ICAG2 | 10.737 | 19 | 1.5218 | .3491 |
| | PG2 | 7.053 | 19 | 1.7472 | .4008 |
| Pair 4 | GCAG1 | 19.944 | 18 | 5.3849 | 1.2692 |
| | PG1 | 14.222 | 18 | 3.4735 | .8187 |
| Pair 5 | GCAG2 | 10.579 | 19 | 1.6095 | .3693 |
| | PG2 | 7.053 | 19 | 1.7472 | .4008 |

**Table 6. SPSS results for Hypothesis 1.**

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | ICAI1 & PI1 | 36 | -.111 | .518 |
| Pair 2 | ICAG1 & PG1 | 36 | -.020 | .909 |
| Pair 3 | ICAG2 & PG2 | 19 | .110 | .654 |
| Pair 4 | GCAG1 & PG1 | 18 | -.028 | .913 |
| Pair 5 | GCAG2 & PG2 | 19 | .186 | .446 |

**Table 7. SPSS results for Hypothesis 1.**

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | ICAI1 - PI1 | 3.5000 | 4.3916 | .7319 | 2.0141 | 4.9859 | 4.782 | 35 | .000 |
| Pair 2 | ICAG1 - PG1 | 5.8611 | 7.1680 | 1.1947 | 3.4358 | 8.2864 | 4.906 | 35 | .000 |
| Pair 3 | ICAG2 - PG2 | 3.6842 | 2.1872 | .5018 | 2.6300 | 4.7384 | 7.342 | 18 | .000 |
| Pair 4 | GCAG1 - PG1 | 5.7222 | 6.4880 | 1.5292 | 2.4958 | 8.9487 | 3.742 | 17 | .002 |
| Pair 5 | GCAG2 - PG2 | 3.5263 | 2.1439 | .4919 | 2.4930 | 4.5597 | 7.169 | 18 | .000 |

## A.3.2 Hypothesis 2

**Table 8. SPSS results for Hypothesis 2.**

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | DIG1 | -5.86 | 36 | 7.168 | 1.195 |
| | DGG1 | -5.72 | 36 | 6.395 | 1.066 |

**Table 9. SPSS results for Hypothesis 2.**

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | DIG1 & DGG1 | 36 | .915 | .000 |

**Table 10. SPSS results for Hypothesis 2.**

**Paired Samples Test**

| | | Paired Differences | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | | | | |
| | | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | t | df | Sig. (2-tailed) |
| Pair 1 | DIG1 - DGG1 | -.139 | 2.900 | .483 | -1.120 | .842 | -.287 | 35 | .776 |

## A.3.3 Hypothesis 3

**Table 11. SPSS results for Hypothesis 3.**

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | DIG2 | -3.68 | 19 | 2.187 | .502 |
| | DGG2 | -3.53 | 19 | 2.144 | .492 |

**Table 12. SPSS results for Hypothesis 3.**

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | DIG2 & DGG2 | 19 | .962 | .000 |

**Table 13. SPSS results for Hypothesis 3.**

**Paired Samples Test**

| | | Paired Differences | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | | | | |
| | | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | t | df | Sig. (2-tailed) |
| Pair 1 | DIG2 - DGG2 | -.158 | .602 | .138 | -.448 | .132 | -1.143 | 18 | .268 |

## A.3.4 Hypothesis 4

**Table 14. SPSS results for Hypothesis 4.**

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | DII1 | -3.50 | 36 | 4.392 | .732 |
| | DIG1 | -5.86 | 36 | 7.168 | 1.195 |

**Table 15. SPSS results for Hypothesis 4.**

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | DII1 & DIG1 | 36 | .764 | .000 |

**Table 16. SPSS results for Hypothesis 4.**

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | | Lower | Upper | | | |
| Pair 1 | DII1 - DIG1 | 2.361 | 4.752 | .792 | .753 | 3.969 | 2.981 | 35 | .005 |

**Table 17. SPSS results for Hypothesis 4.**

**Group Statistics**

| | Grouping variable A | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| There is no difference in accuracy for individual assessments between tasks (G1, G2). | DIG1 | 28 | -7.54 | 7.037 | 1.330 |
| | DIG2 | 20 | -3.90 | 2.337 | .523 |

**Table 18. SPSS results for Hypothesis 4.**

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| There is no difference in accuracy for individual assessments between tasks (G1, G2). | Equal variances assumed | 12.638 | .001 | -2.219 | 46 | .031 | -3.636 | 1.638 | -6.934 | -.338 |
| | Equal variances not assumed | | | -2.544 | 34.805 | .016 | -3.636 | 1.429 | -6.537 | -.734 |

**Table 19. SPSS results for Hypothesis 4.**

**Group Statistics**

| | Grouping variable B | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| There is no difference in accuracy for individual assessments between tasks (I1, G2). | DII1 | 36 | -3.50 | 4.392 | .732 |
| | DIG2 | 20 | -3.90 | 2.337 | .523 |

**Table 20. SPSS results for Hypothesis 4.**

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| There is no difference in accuracy for individual assessments between tasks (I1, G2). | Equal variances assumed | 5.007 | .029 | .378 | 54 | .707 | .400 | 1.059 | -1.723 | 2.523 |
| | Equal variances not assumed | | | .445 | 53.953 | .658 | .400 | .899 | -1.403 | 2.203 |

## A.3.5 Hypothesis 5

**Table 21. SPSS results for Hypothesis 5.**

**Group Statistics**

| | Grouping variable | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| There is no difference in accuracy for pair assessments between tasks (G1, G2) | DGG1 | 14 | -7.21 | 6.495 | 1.736 |
| | DGG2 | 19 | -3.53 | 2.144 | .492 |

**Table 22. SPSS results for Hypothesis 5.**

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| There is no difference in accuracy for pair assessments between tasks (G1, G2) | Equal variances assumed | 9.842 | .004 | -2.321 | 31 | .027 | -3.688 | 1.589 | -6.929 | -.447 |
| | Equal variances not assumed | | | -2.044 | 15.101 | .059 | -3.688 | 1.804 | -7.531 | .155 |

## A.3.6 Hypothesis 6

**Table 23. SPSS results for Hypothesis 6.**

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | ICAG1 | 20.083 | 36 | 6.2307 | 1.0384 |
| | GCAG1 | 19.944 | 36 | 5.3074 | .8846 |

**Table 24. SPSS results for Hypothesis 6.**

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | ICAG1 & GCAG1 | 36 | .886 | .000 |

**Table 25. SPSS results for Hypothesis 6.**

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | | Lower | Upper | | | |
| Pair 1 | ICAG1 - GCAG1 | .1389 | 2.8998 | .4833 | -.8423 | 1.1200 | .287 | 35 | .776 |

## A.3.7 Hypothesis 7

**Table 26. SPSS results for Hypothesis 7.**

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | ICAG2 | 10.600 | 20 | 1.6026 | .3584 |
| | GCAG2 | 10.500 | 20 | 1.6059 | .3591 |

**Table 27. SPSS results for Hypothesis 7.**

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | ICAG2 & GCAG2 | 20 | .920 | .000 |

**Table 28. SPSS results for Hypothesis 7.**

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | | Lower | Upper | | | |
| Pair 1 | ICAG2 - GCAG2 | .1000 | .6407 | .1433 | -.1999 | .3999 | .698 | 19 | .494 |

## A.3.8 Hypothesis 8

**Table 29. SPSS results for Hypothesis 8.**

**Group Statistics**

| | Grouping variable | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| There is no difference between performances of task I1 and G2. | PI1 | 27 | 7.259 | 2.9299 | .5639 |
| | PG2 | 19 | 7.053 | 1.7472 | .4008 |

**Table 30. SPSS results for Hypothesis 8.**

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 95% Confidence Interval of the Difference | |
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | Lower | Upper |
| There is no difference between performances of task I1 and G2. | Equal variances assumed | 5.719 | .021 | .274 | 44 | .785 | .2066 | .7529 | -1.3107 | 1.7239 |
| | Equal variances not assumed | | | .299 | 43.041 | .767 | .2066 | .6918 | -1.1885 | 1.6017 |

## A.4 Questionnaire results

In this Section are the results from the questionnaire presented.

### A.4.1 Question 1 –

Have you ever come across the game Mastermind before? Yes No

If **Yes**, please describe your experience briefly.

**Table 31. SPSS results for Question 1.**

**Question1**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Did not understand the question | 3 | 6.3 | 6.3 | 6.3 |
| | Have seen other people playing it | 1 | 2.1 | 2.1 | 8.3 |
| | No | 20 | 41.7 | 41.7 | 50.0 |
| | Played it recently | 4 | 8.3 | 8.3 | 58.3 |
| | Played similar games | 1 | 2.1 | 2.1 | 60.4 |
| | Played the game at friends' places as a child | 2 | 4.2 | 4.2 | 64.6 |
| | Played the game at home as a child | 17 | 35.4 | 35.4 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

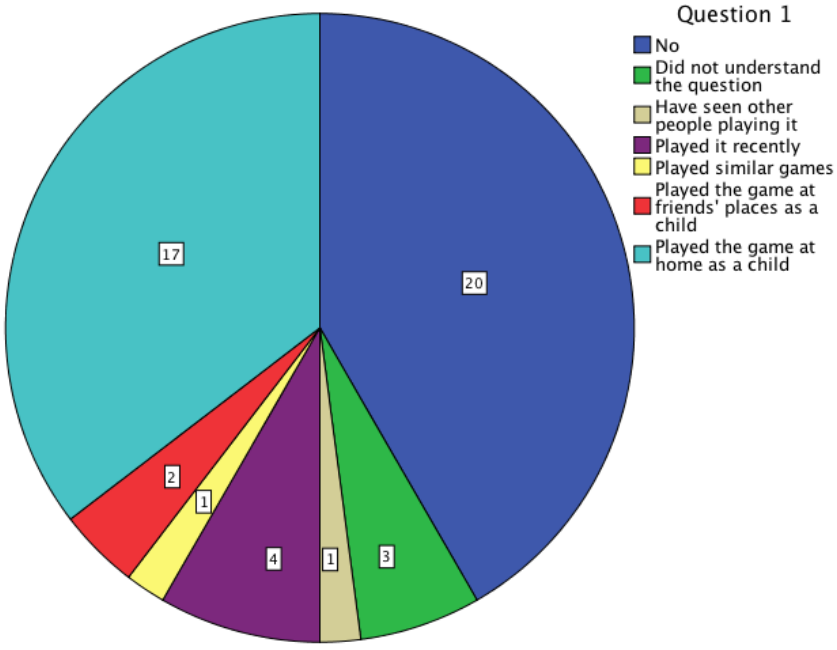



**Figure 23. Response distribution for Question 1.**

### A.4.2 Question 2 –

Have you ever come across capability assessments before? Yes No

If **Yes**, in which context?

**Table 32. SPSS results for Question 2.**

**Question 2**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Did not understand the question | 1 | 2.1 | 2.1 | 2.1 |
| | No | 33 | 68.8 | 68.8 | 70.8 |
| | Through military service | 2 | 4.2 | 4.2 | 75.0 |
| | Through my occupation | 4 | 8.3 | 8.3 | 83.3 |
| | Through sports | 2 | 4.2 | 4.2 | 87.5 |
| | Through studies (University) | 6 | 12.5 | 12.5 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

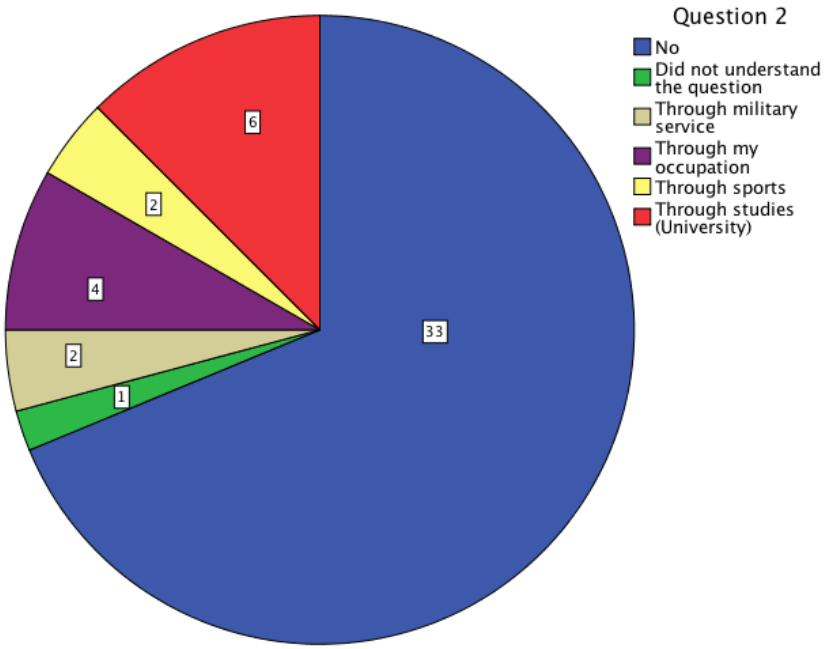
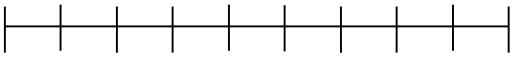



**Figure 24. Response distribution for Question 2.**

### A.4.3 Question 3 –

Did you experience it harder to do an individual assessment of a group task or to do a group assessment of a group task? Please put a cross on the line below where appropriate.

Individual assessment much harder.  Group assessment much harder.

If a specific task was harder, please explain why:

**Table 33. SPSS results for Question 3.**

**Statistics**

3a

| | | |
|---|---|---|
| N | Valid | 48 |
| | Missing | 0 |
| Mean | | .58 |
| Median | | .00 |
| Minimum | | -2 |
| Maximum | | 5 |

**Table 34. SPSS results for Question 3.**

**3a**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | -2 | 7 | 14.6 | 14.6 | 14.6 |
| | -1 | 10 | 20.8 | 20.8 | 35.4 |
| | 0 | 8 | 16.7 | 16.7 | 52.1 |
| | 1 | 6 | 12.5 | 12.5 | 64.6 |
| | 2 | 8 | 16.7 | 16.7 | 81.3 |
| | 3 | 7 | 14.6 | 14.6 | 95.8 |
| | 4 | 1 | 2.1 | 2.1 | 97.9 |
| | 5 | 1 | 2.1 | 2.1 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

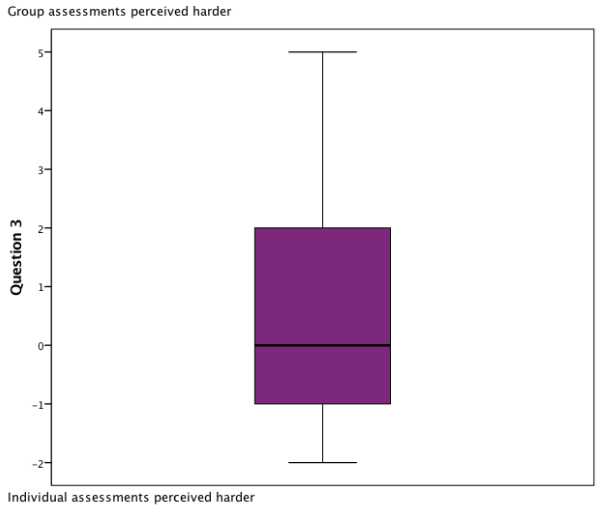



**Figure 25. Response distribution for Question 3.**

**Table 35. SPSS results for Question 3.**

**Question 3**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Both were equally difficult | 3 | 6.3 | 6.3 | 6.3 |
| | Chose not to respond | 13 | 27.1 | 27.1 | 33.3 |
| | Did not understand the question | 12 | 25.0 | 25.0 | 58.3 |
| | More difficult to assess in group as it is then required to understand the other person's view | 7 | 14.6 | 14.6 | 72.9 |
| | More difficult to assess individually as the other person's capability is unknown | 13 | 27.1 | 27.1 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

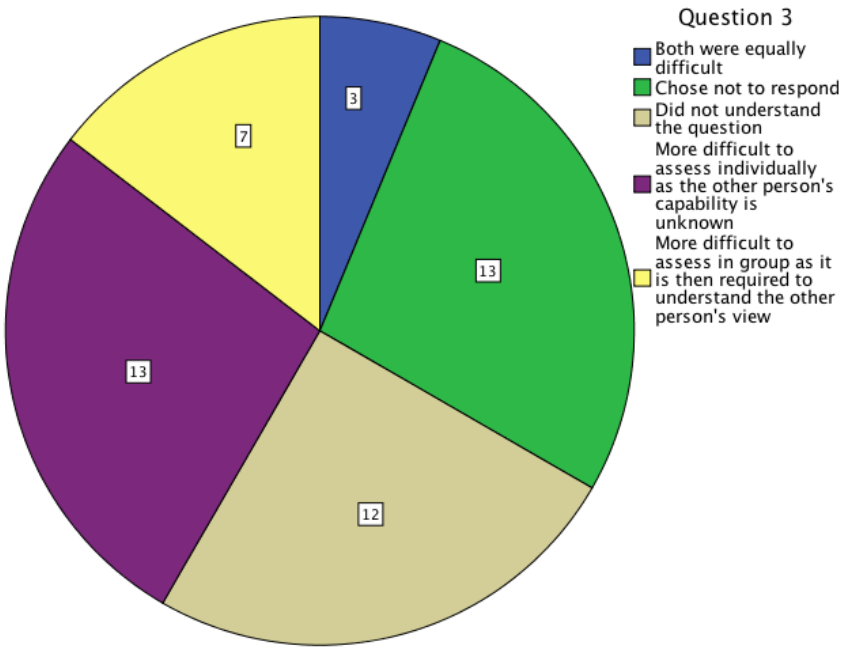



**Figure 26. Response distribution for Question 3.**

### A.4.4 Question 4 –

Did you think differently when reasoning in the individual assessments than in the group assessments? Yes No

If **Yes**, how did it differ? If **No**, how did you reason?

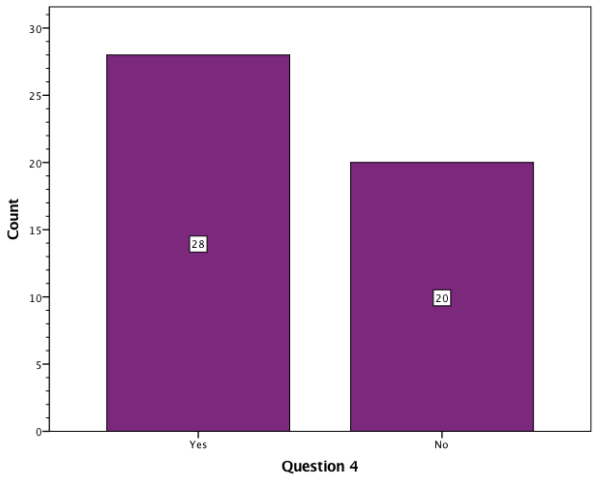



**Figure 27. Response distribution for Question 4.**

**Table 36. SPSS results for Question 4.**

**Question 4**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Chose not to respond | 8 | 16.7 | 16.7 | 16.7 |
| | Did not understand the question | 15 | 31.3 | 31.3 | 47.9 |
| | No, assumed the mean from practice period and weighted the two when in pairs | 1 | 2.1 | 2.1 | 50.0 |
| | No, chance of solving the code does not depend on the amount of actors | 2 | 4.2 | 4.2 | 54.2 |
| | Yes, assumed the mean from practice period and weighted the two when in pairs | 6 | 12.5 | 12.5 | 66.7 |
| | Yes, assumed the pair was stronger than each individual | 8 | 16.7 | 16.7 | 83.3 |
| | Yes, assumed the pair was weaker than each individual | 8 | 16.7 | 16.7 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

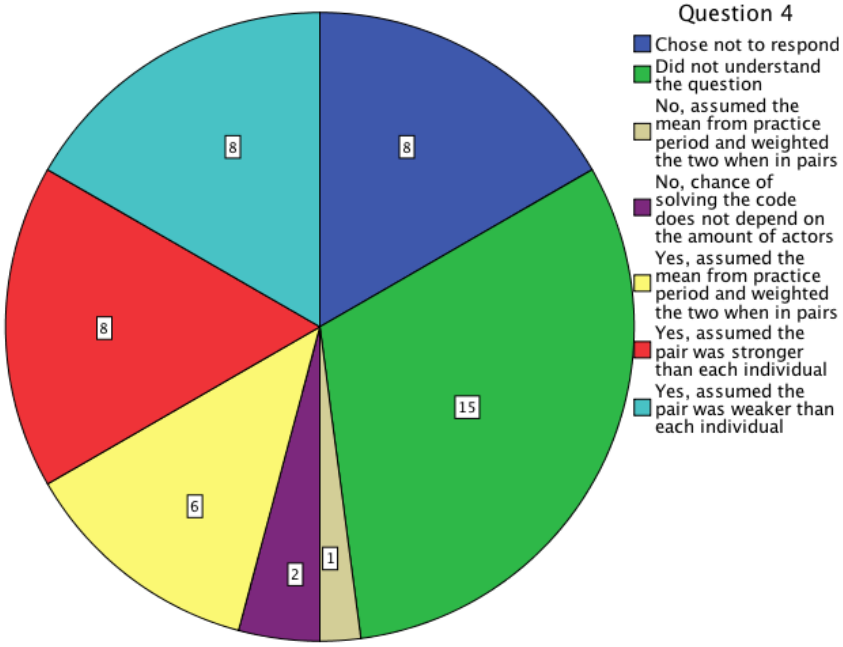



**Figure 28. Response distribution for Question 4.**

### A.4.5 Question 5 –

Did you think differently in your individual assessments between the individual tasks and the group tasks? Yes No

If **Yes**, how did it differ? If **No**, how did you reason?

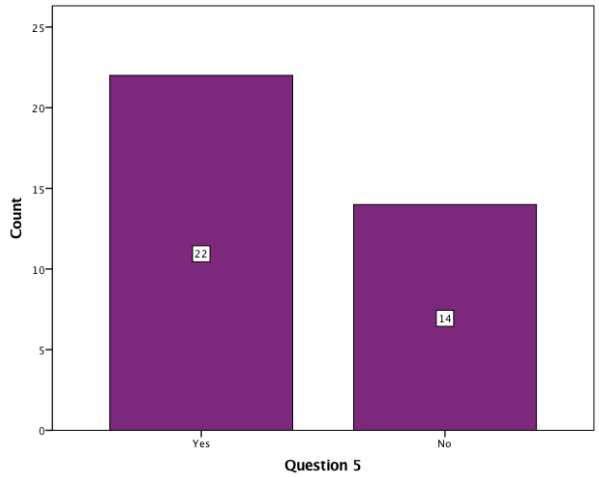



**Figure 29. Response distribution for Question 5.**

**Table 37. SPSS results for Question 5.**

**Question 5**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Chose not to respond | 5 | 10.4 | 10.4 | 10.4 |
| | Did not understand the question | 7 | 14.6 | 14.6 | 25.0 |
| | No, assumed means from practice period and applied to pair | 3 | 6.3 | 6.3 | 31.3 |
| | No, the pair seemed to be equal | 4 | 8.3 | 8.3 | 39.6 |
| | Not applicable | 11 | 22.9 | 22.9 | 62.5 |
| | Yes, assumed means from practice period and applied to pair | 1 | 2.1 | 2.1 | 64.6 |
| | Yes, assumed the pair was stronger than each indvidual | 3 | 6.3 | 6.3 | 70.8 |
| | Yes, assumed the pair was weaker than each indvidual | 14 | 29.2 | 29.2 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

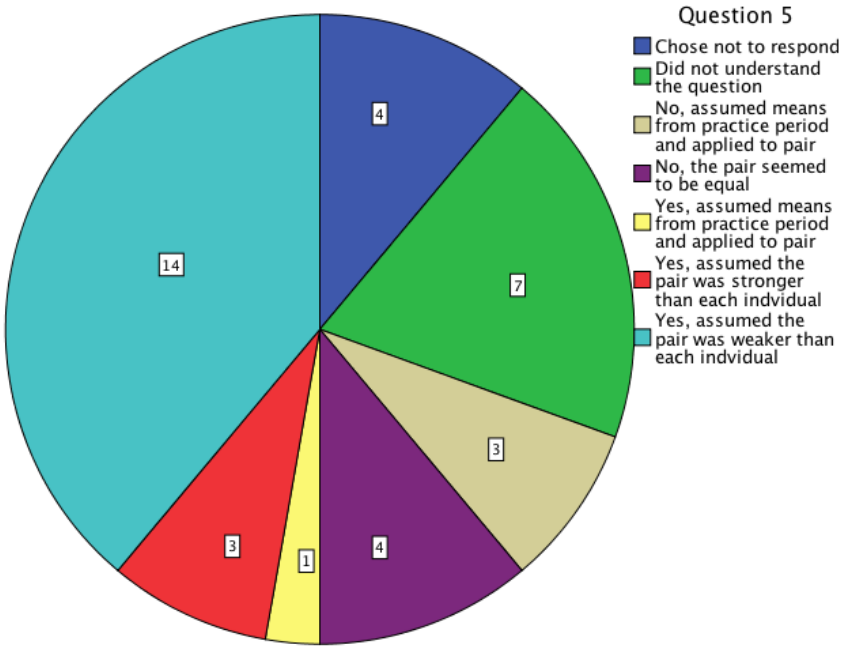



**Figure 30. Response distribution for Question 5.**

### A.4.6 Question 6 –

Did you experience it harder to do the assessments of the first group task or the assessments of the second group task? Please put a cross on the line below where appropriate.

First group task (G1) much harder. ——|——|——|——|——|——|——|——|——| Second group task (G2) much harder.

If a specific task was harder, please explain why:

**Table 38. SPSS results for Question 6.**

**Statistics**

6a

| | | |
|---|---|---|
| N | Valid | 8 |
| | Missing | 40 |
| Mean | | 1.00 |
| Median | | 2.50 |
| Minimum | | -4 |
| Maximum | | 4 |

**Table 39. SPSS results for Question 6.**

**Question 6**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | -4 | 1 | 2.1 | 12.5 | 12.5 |
| | -3 | 1 | 2.1 | 12.5 | 25.0 |
| | -1 | 1 | 2.1 | 12.5 | 37.5 |
| | 2 | 1 | 2.1 | 12.5 | 50.0 |
| | 3 | 2 | 4.2 | 25.0 | 75.0 |
| | 4 | 2 | 4.2 | 25.0 | 100.0 |
| | Total | 8 | 16.7 | 100.0 | |
| Missing | 999 | 40 | 83.3 | | |
| Total | | 48 | 100.0 | | |

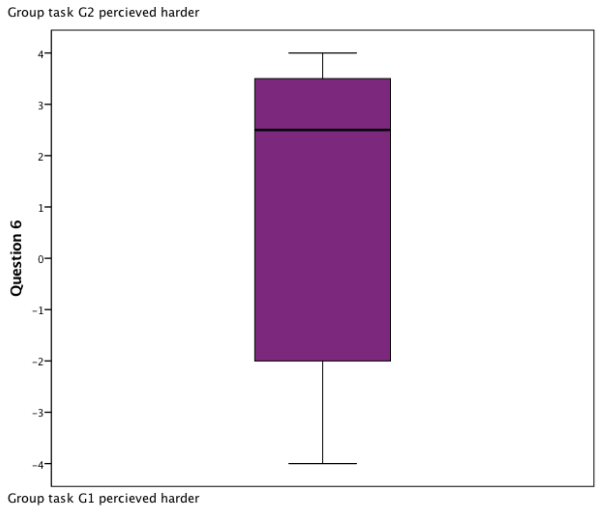



**Figure 31. Response distribution for Question 6.**

**Table 40. SPSS results for Question 6.**

**Question 6**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Chose not to respond | 1 | 2.1 | 2.1 | 2.1 |
| | Depends on experience, got to know the other person in the pair | 1 | 2.1 | 2.1 | 4.2 |
| | Did not understand the question | 2 | 4.2 | 4.2 | 8.3 |
| | Less opportunity to affect the result in G1 | 1 | 2.1 | 2.1 | 10.4 |
| | More difficult to continue someone else's work (G2) | 2 | 4.2 | 4.2 | 14.6 |
| | More room for errors (G1) | 1 | 2.1 | 2.1 | 16.7 |
| | Not applicable | 40 | 83.3 | 83.3 | 100.0 |
| | Total | 48 | 100.0 | 100.0 | |

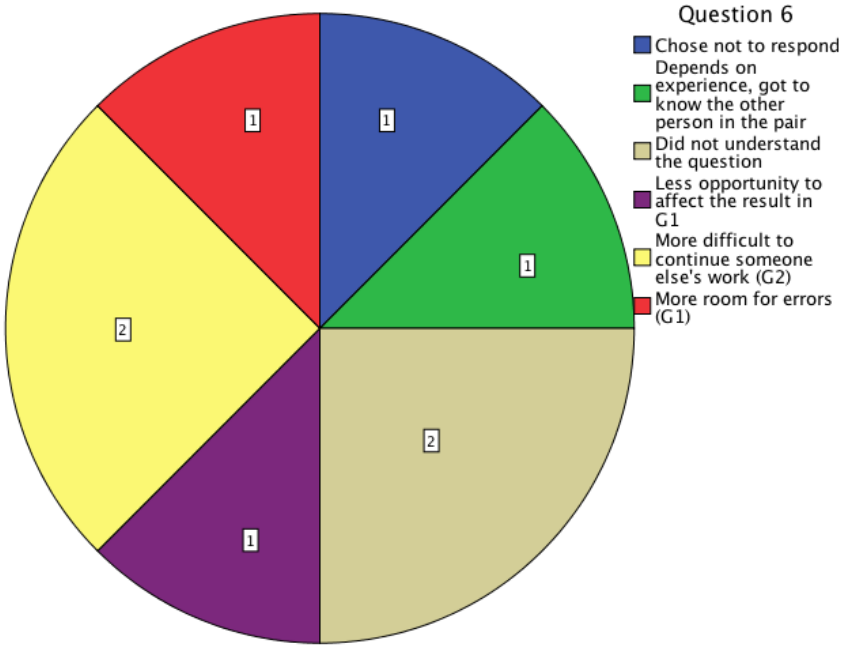



**Figure 32. Response distribution for Question 6.**

### A.4.7 Question 7 –

Other comments regarding the experiment (E.g level of difficulty of tasks, was it easy to follow the instructions and so on).

Question 7 generated the following categorised comments:

- Good and clear instructions
- Good level of difficulty
- Questionnaire could have been easier to understand
- Underestimated my capability
- Good with a practice session where questions could be asked before the test started
- Difficult to understand which strategy the partner used, when the strategy was figured out it was easier to solve the code.
- It is important with a safety margin. It feels safer to choose a row number which is higher then the most probable row, than gamble on a good assessment and fail.
- Probably making the task harder through overanalysing it.

## A.5 Confidence in capability assessments

In this Section are the results from the statistical analyses made in SPSS presented.

### A.5.1 Individual capability assessments for individual tasks I1.

**Table 41. Confidence in assessments related to ICAI1.**

**Statistics**

| | | ICAI1 +-0 | ICAI1 +-1 | ICAI1 +-2 |
|---|---|---|---|---|
| N | Valid | 28 | 28 | 28 |
| | Missing | 62 | 62 | 62 |
| Mean | | 4.25 | 5.61 | 7.36 |
| Median | | 4.00 | 5.50 | 8.00 |
| Range | | 9 | 8 | 7 |
| Minimum | | 0 | 2 | 3 |
| Maximum | | 9 | 10 | 10 |
| Percentiles | 25 | 3.00 | 4.00 | 6.25 |
| | 50 | 4.00 | 5.50 | 8.00 |
| | 75 | 6.00 | 6.75 | 9.00 |

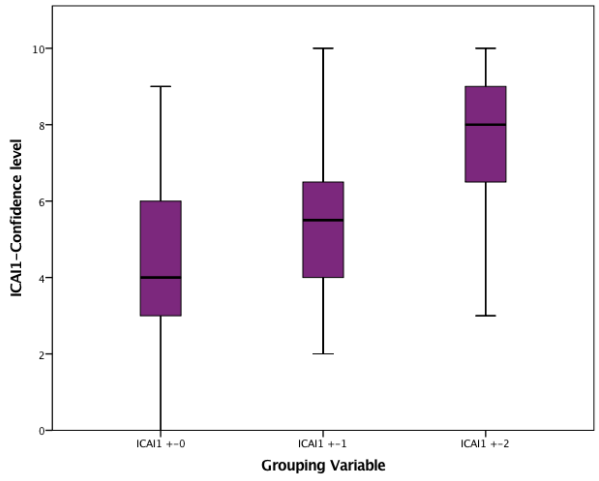



**Figure 33. Response distribution for ICAI1.**

## A.5.2 Individual capability assessments for group tasks G1.

**Table 42. Confidence in assessments related to ICAG1.**

**Statistics**

| | | ICAG1 +-0 | ICAG1 +-1 | ICAG1 +-2 |
|---|---|---|---|---|
| N | Valid | 30 | 30 | 30 |
| | Missing | 60 | 60 | 60 |
| Mean | | 3.43 | 4.83 | 6.57 |
| Median | | 3.00 | 5.00 | 6.50 |
| Range | | 7 | 6 | 8 |
| Minimum | | 0 | 1 | 2 |
| Maximum | | 7 | 7 | 10 |
| Percentiles | 25 | 2.00 | 3.00 | 5.00 |
| | 50 | 3.00 | 5.00 | 6.50 |
| | 75 | 5.00 | 6.25 | 8.00 |

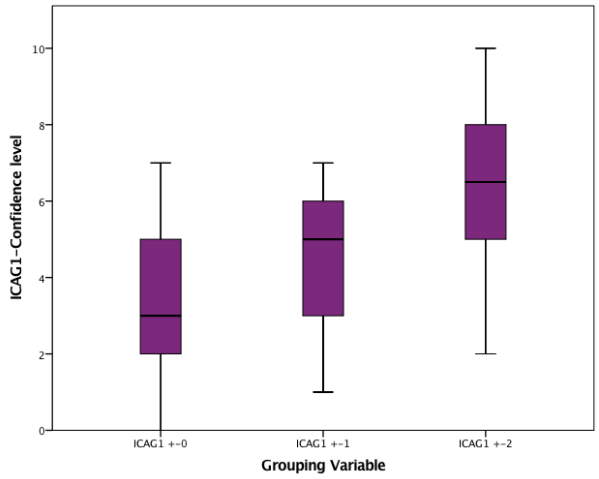



**Figure 34. Response distribution for ICAG1.**

## A.5.3 Paired capability assessments for group tasks G1.

**Table 43. Confidence in assessments related to GCAG1.**

**Statistics**

| | | GCAG1 +-0 | GCAG1 +-1 | GCAG1 +-2 |
|---|---|---|---|---|
| N | Valid | 17 | 17 | 17 |
| | Missing | 73 | 73 | 73 |
| Mean | | 3.41 | 4.88 | 6.71 |
| Median | | 3.00 | 4.00 | 7.00 |
| Range | | 4 | 4 | 5 |
| Minimum | | 2 | 3 | 4 |
| Maximum | | 6 | 7 | 9 |
| Percentiles | 25 | 2.00 | 4.00 | 5.00 |
| | 50 | 3.00 | 4.00 | 7.00 |
| | 75 | 5.00 | 6.00 | 8.00 |

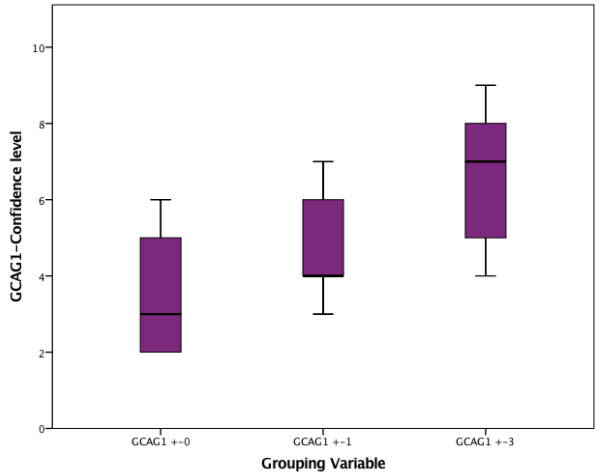



**Figure 35. Response distribution for GCAG1.**

## A.5.4 Individual capability assessments for group tasks G2.

**Table 44. Confidence in assessments related to ICAG2.**

**Statistics**

| | | ICAG2 +-0 | ICAG2 +-1 | ICAG2 +-2 |
|---|---|---|---|---|
| N | Valid | 19 | 19 | 19 |
| | Missing | 71 | 71 | 71 |
| Mean | | 4.05 | 5.53 | 6.79 |
| Median | | 4.00 | 6.00 | 8.00 |
| Range | | 6 | 6 | 6 |
| Minimum | | 1 | 2 | 4 |
| Maximum | | 7 | 8 | 10 |
| Percentiles | 25 | 3.00 | 4.00 | 4.00 |
| | 50 | 4.00 | 6.00 | 8.00 |
| | 75 | 5.00 | 7.00 | 9.00 |

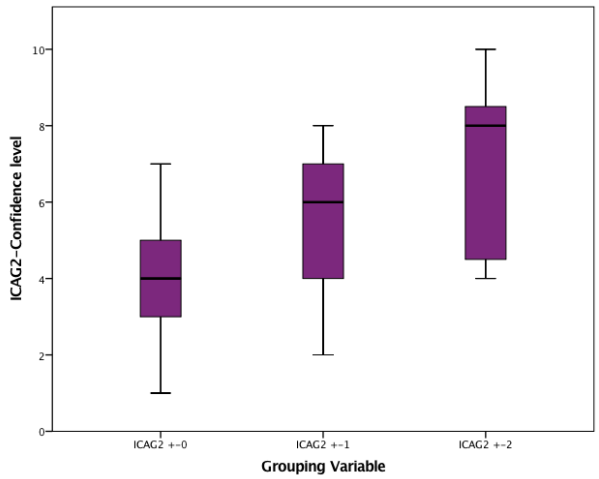



**Figure 36. Response distribution for ICAG2.**

## A.5.5 Paired capability assessments for group tasks G2.

**Table 45. Confidence in assessments related to GCAG2.**

**Statistics**

| | | GCAG2 +-0 | GCAG2 +-1 | GCAG2 +-2 |
|---|---|---|---|---|
| N | Valid | 10 | 10 | 10 |
| | Missing | 80 | 80 | 80 |
| Mean | | 4.30 | 6.00 | 7.30 |
| Median | | 4.00 | 6.00 | 7.50 |
| Range | | 4 | 4 | 4 |
| Minimum | | 3 | 4 | 5 |
| Maximum | | 7 | 8 | 9 |
| Percentiles | 25 | 3.00 | 4.75 | 6.00 |
| | 50 | 4.00 | 6.00 | 7.50 |
| | 75 | 5.25 | 7.25 | 8.25 |

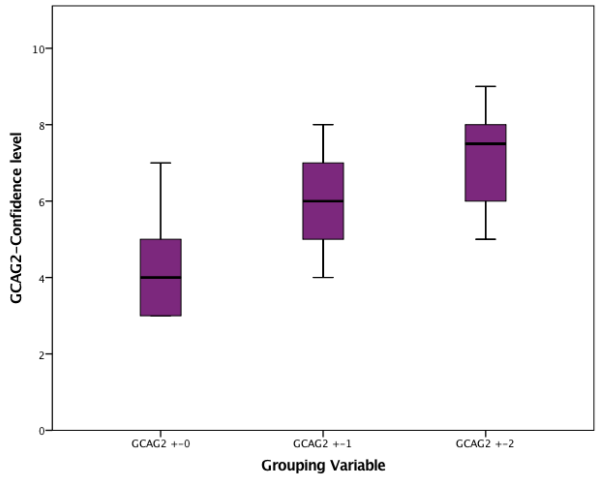



**Figure 37. Response distribution for GCAG2.**

# Appendix B – Instructions

In this Appendix the instructions to the participants are presented. The instructions were used as a manuscript for the supervisors to follow.

## B.1 Instructions in Swedish

### Inledande information

Experimentet är del av Sebastian Severinsens och Malin Hansons examensarbete inom ramen för examensarbete vid avdelningen för riskhantering och samhällssäkerhet vid Lunds universitet. Examensarbetet är en del i det pågående forskningsprojektet om förmågebedömningar (PRIVAD, Programme for Risk and Vulnerability Analysis Development) som tar avstamp i det svenska krishanteringssystemet och de förmågebedömningar som genomförs i samband med risk och sårbarhetsanalyser.

### Introduktion till Mastermind

Spelet som ni kommer att spela heter Mastermind och ser ut så här [visa spelplan]. Spelet går ut på att lösa en kod på fyra variabler [peka], varje variabel kan vara en av sex olika färger. Alla färger kan användas 0-4 gånger. Ni börjar med att gissa en kombination av färger och experimentledaren kommer sedan att markera på spelplanen om ni har gissat rätt eller fel här [peka]. Svart betyder rätt färg på rätt plats, kryss betyder rätt färg på fel plats och vit fel färg. Feedbacken ges i ordningen svart, kryss och vit eftersom varje feedbackruta inte motsvarar gissningsrutorna. Efter ett par rader kommer ni att kunna börja dra slutsatser om vilken/vilka färger som ska vara med och var de ska vara. Observera att ni inte kommer att kunna ”sudda”, utan vald färg ligger.

Innan varje uppgift kommer ni få göra en förmågebedömning där ni noterar den rad ni bedömer att ni kommer att lösa koden på samt hur säkra ni är på den bedömningen. Tänk på när ni gör er bedömning att ni är i den här situationen, ni kanske är nybörjare på spelet, det kommer finnas ljud omkring, viss tidspress kan finnas samt andra deltagare kan titta på. En vanlig spelplan brukar ha mellan 8-12 rader, men ni har så många rader på er som ni vill. Innan vi börjar kommer ni få öva under 10 minuter. Försök göra det mesta av tiden genom att ha ett bra tempo i era gissningar, så ni får möjlighet att klura på hur spelet fungerar. Det är nämligen det som är syftet med övningen, att ni ska förstå hur spelet fungerar. Nu parar vi ihop er två och två. Varje par kommer ha en experimentledare under experimentets gång.

---

### G1

Nu får ni er första uppgift. Under 10 minuter har ni nu övat och har lite bättre koll på hur spelet fungerar. I den här förmågebedömningen ni ska göra, skriv ner på vilken rad ni tror att ni kommer att lösa uppgiften på, samt hur säkra ni är på det. [Dela ut papper]

I nästa förmågebedömning ska ni bedöma vid vilken rad ni kommer att lösa uppgiften när ni räknar med även er partners rader (det vill säga antalet sammanlagda rader, tex. 7+9=16). Båda två kommer ha varsin spelplan och lösa varsin uppgift. Ingen kommunikation får lov att utföras. Ni har ingen tidsgräns, men det är rekommenderat att hålla ett bra tempo under uppgiften. Observera att ni gör den här bedömningen själv. [Dela ut papper]

I sista förmågebedömning ska ni bedöma tillsammans vid vilken rad ni kommer att lösa uppgiften när era resultat läggs ihop. Observera att ni måste komma fram till ett gemensamt svar. [Dela ut papper]

Har ni några frågor?

Varsågod att börja.

---

### G2

Här ska ni lösa en omgång tillsammans. En av er kommer att börja lägga 4 rader och sedan tar den andra över. Ingen kommunikation får lov att göras och den som inte börjar får vända sig bort under tiden den väntar på sin tur. Ni har ingen tidsgräns, men det är rekommenderat att ha ett bra tempo under uppgiften. Ni ska nu göra en individuell bedömning om vid vilken rad ni tror att ni gemensamt kommer att lösa uppgiften.

[Dela ut papper]

Ni ska nu gemensamt bedöma vid vilken rad ni tror att ni kommer lösa uppgiften. Om ni löser problemet på 4 rader eller mindre så kommer ni få göra om uppgiften, nu med den andra deltagaren först.

[Dela ut papper]

Har ni några frågor?

Varsågod att börja.

---

### Enkät

Tack för er medverkan, det kommer att komma en enkät på er email. Vi är väldigt tacksamma om ni vill fylla i den eftersom den är en del av experimentet. Om ni vill fylla i enkäten här så har vi pappersenkäter.

## B.2 Instructions in English

### Project introduction

This experiment is part of Sebastian Severinsen's and Malin Hanson's degree thesis for the Division of Risk Management and Societal Safety at Lund University. The degree thesis is part of the PRIVAD (Programme for Risk and Vulnerability Analysis Development) project, in which the concept of capability assessments in risk and vulnerability analysis is further explored.

### Introduction to Mastermind

The game you'll play is called Mastermind and looks like this [show game board]. To win the game, you are to solve a code of four dots [point], each dot can be one out of six colours. All colours can be used 0-4 times. You start by guessing a combination of colours and the experiment supervisor will give you feedback on your guess [show dots]. Black means right colour in the right position, cross means right colour but in the wrong position, and white means wrong colour. The order of the feedback is black, cross and white as the feedback dots doesn't represent a "guess dot". After a couple of rows you'll be able to draw conclusions of which colours are in the code and where they should be located.

Before each task, you'll do a capability assessment where you assess at which row you'll solve the code. You'll also assess how certain you are of your assessment. Keep in mind that this is the situation you will find whilst solving the code, people are watching, there will be ambient sound and so on. This is something to take into consideration when you assess the capability. A regular game board has between 8-12 rows, but you'll be able to use as many rows as you like. Before we begin you'll practise for 10 minutes. Try to make the most of the time you got by keeping a good pace throughout your guesses, which allows you to figure out the game. This is the purpose of exercising, to understand how the game works. Now we'll team you up in groups of two. Each group will have an experiment supervisor.

---

### G1

Now you will get your first assignment. After practising for 10 minutes you hopefully have a better understanding of the game and how it works. In this capability assessment you will write down at which row you believe you will be able to solve the code and how certain you are on your assessment. [Distribute paper]

Next capability assessment is aimed towards what row you think you will solve the code when you take your partners result into account. In other words you will assess the total amount of rows you and your partner will use to solve one code each. For example 7+9=16. No communication will be allowed and there is no time limit. However, we recommend that you keep a good pace throughout the task. The assessment will be conducted individually

The last capability assessment will be conducted in cooperation with your partner and you will together assess at which row you will solve two codes. Note that you have to reach a mutual response. [Distribute paper]

Any questions?

You may begin.

### G2

Here you'll solve the code together. One of you will do the first four rows and then the second person will continue. You have no time limit, but it's recommended to keep a good pace while solving the code. No communication is allowed and the second person will face the other way while waiting on their turn. Before the task you'll make an individual assessment of which row you think your group will solve the task.

[Hand out paper.]

Now you'll discuss and together estimate on which row you think you'll solve the game.

[Hand out paper.]

Any questions?

You may begin.

---

### Questionnaire

Thank you for your contribution. A questionnaire will be sent to your email, please fill it in as it's part of the experiment. If you'd prefer to fill it here on the spot, we have a few copies up front.

# Appendix C – Capability assessment form

Fylls i av experimentledaren/ To be used by experiment supervisor

I1 / G1 / G2

Individuell bedömning / Gruppbedömning

KOD:_____

## Förmågebedömning

**Vilken rad bedömer du att uppgiften kommer lösas på? Rad _____**

**Hur säker är du på din bedömning:**
På en skala från 0-10, markera ditt svar.

Hur säker är du/ni på att uppgiften löses på den bedömda raden?
Inte säker – 0 1 2 3 4 5 6 7 8 9 10 – Helt säker

Hur säker är du/ni på att uppgiften löses inom ett intervall från den bedömda raden på +/-1 rad?
Inte säker – 0 1 2 3 4 5 6 7 8 9 10 – Helt säker

Hur säker är du/ni på att uppgiften löses inom ett intervall från den bedömda raden på +/-2 rader?
Inte säker – 0 1 2 3 4 5 6 7 8 9 10 – Helt säker

---

## Capability assessment

**At what row do you assess the task will be solved? Row _____**

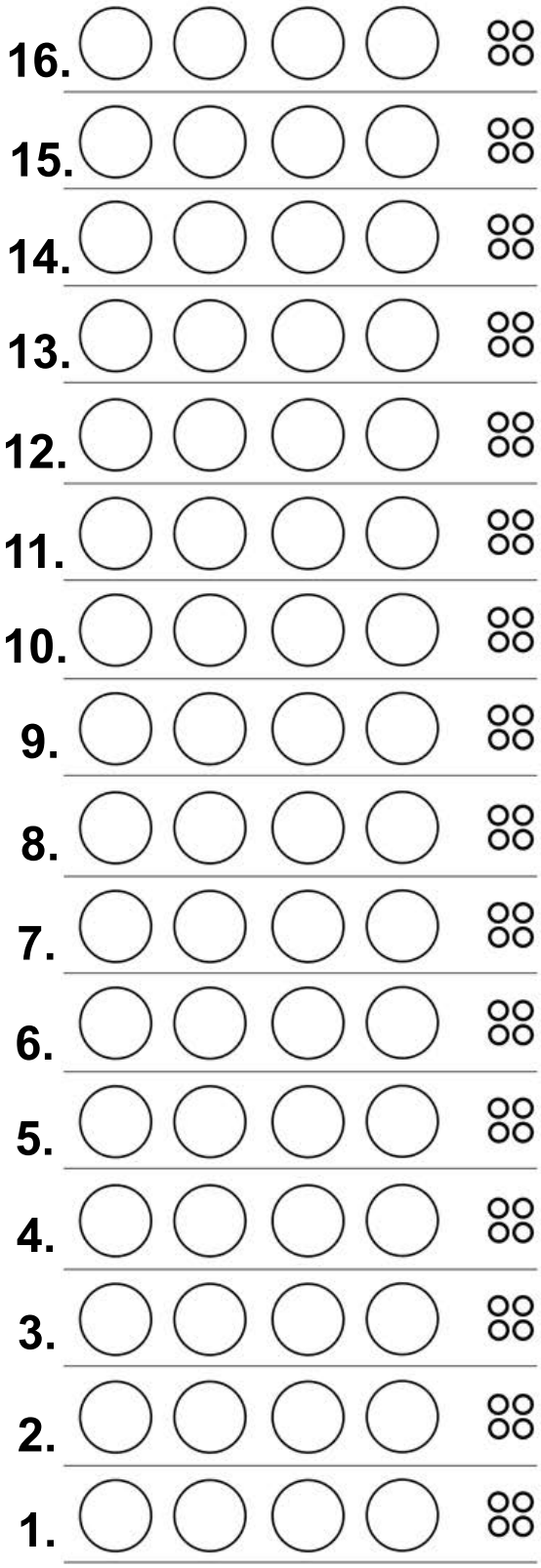
**How confident are you in your assessment:**
On a scale from 0-10, please mark your answer.

How confident are you that the task will be solved on the expected row?
Not confident – 0 1 2 3 4 5 6 7 8 9 10 – Completely confident

How confident are you that the task will be solved within an interval of +/-1 row of the expected row?
Not confident – 0 1 2 3 4 5 6 7 8 9 10 – Completely confident

How confident are you that the task will be solved within an interval of +/-2 rows of the expected row?
Not confident – 0 1 2 3 4 5 6 7 8 9 10 – Completely confident

## Appendix D – Mastermind Game board




# Spelregler i korthet:

Gissa vilka fyra färger koden består av.

- Färger att välja på:
  Blå Gul Grön Röd Orange Svart
- Varje färg kan användas 0-4 gånger
- Feedback ges inte ifrån experimentledaren förrän hela raden är ifylld.
- Feedback färger:
  Svart - Rätt färg, rätt plats
  Kryss - Rätt färg, fel plats
  Vit - Fel färg
- Feedbackrutorna representerar inte en specifik gissningsruta

# Game rules in short:

Guess which four colours the code is made of.

- Colours to choose from:
  Blue Yellow Green Red Orange Black.
- Every colour may be repeated 0-4 times
- Feedback is not given by experiment supervisor until a row is compleated.
- Feedback colours:
  Black - Right colour, right position
  Cross - Right colour, wrong position
  White - Wrong colour
- A feedback box does not represent a specific guess box.

Fylls i av experimentledaren/
To be used by experiment supervisor

I1 / G1 / G2
KOD: _____

# Appendix E – Abbreviations

During the data collection and the statistical analyses, it was required to adopt abbreviations. The following naming system was used:

WXYZ, where:

W = D, difference between capability assessment and performance. Excluded where CA or P is used.
X = I or G, describes if the task or assessment was made individually (I) or in pairs (G).
X = CA or P, which describes if it is a capability assessment (CA) or the actual performance (P). Excluded where D is used.
Y= I1, G1 or G2, which describes which task it concerns, the individual task (I1), the type 1 dependency (G1), or the type 2 dependency (G2).

The abbreviations used are as follows:

| | |
|---|---|
| ICAI1 | Individual Capability Assessment for I1 |
| ICAG1 | Individual Capability Assessment for G1 |
| ICAG2 | Individual Capability Assessment for G2 |
| GCAG1 | Pair Capability Assessment for G1 |
| GCAG2 | Pair Capability Assessment for G2 |
| DII1 | Difference between Individual Capability Assessment and Performance for I1 |
| DIG1 | Difference between Individual Capability Assessment and Performance for G1 |
| DIG2 | Difference between Individual Capability Assessment and Performance for G2 |
| DGG1 | Difference between Pair Capability Assessment and Performance for G2 |
| DGG2 | Difference between Pair Capability Assessment and Performance for G2 |
| PI1 | Performance for I1 |
| PG1 | Performance for G1 |
| PG2 | Performance for G2 |