How well do capability assessments reflect actual capability? – An experimental study of capability assessments with multi-actor dependencies

Hanson, Malin; Severinsen, Sebastian; Lindbom, Hanna

Published in:
Risk, Reliability and Safety: Innovating Theory and Practice

DOI:
10.1201/9781315374987-70

2016

Document Version:
Peer reviewed version (aka post-print)

How well do capability assessments reflect actual capability? – An experimental study of capability assessments with multi-actor dependencies

M. Hanson, S. Severinsen & H. Lindbom
Division of Risk Management and Societal Safety, Lund University, Sweden
Centre for Critical Infrastructure Protection Research (CenCIP), Lund University, Sweden
Lund University Centre for Risk Assessment and Management, Sweden
Centre for Societal Resilience, Lund University, Sweden

ABSTRACT: In order to cater for the increasingly complex society several countries have adopted a capabilities-based planning approach, including capability assessments, as part of their preparedness for disasters. Since the number of actors who are dependent on each other is increasing, capability assessments are made based on assumptions of other actors' capabilities. It is therefore crucial to understand how multi-actor dependencies affect capability assessments. This experimental study aims at investigating how well capability assessments correspond to actual capability, taking multi-actor dependencies into account. A total of 48 participants, randomly assigned into 24 pairs, assessed their individual capability and capability as a group to accomplish a given task. The results show that the participants in general underestimated their capability, both individually and in groups.

1 INTRODUCTION

Capability assessments and capabilities-based planning are becoming increasingly popular as part of the preparedness for disasters in order to prepare for a wide variety of risks and threats instead of specific scenarios (Programme National Security 2007). Several countries, including Australia, The Netherlands, New Zealand, Sweden, The UK and The USA now use capability assessments as part of their emergency preparedness (Australian Capital Territory 2012, Cabinet Office 2014, Dutch Ministry of Interior and Kingdom Relations 2009, Homeland Security 2013, Houdijk 2010, Ministry of Civil Defence and Emergency Management n.d., Swedish Civil Contingencies Agency 2014). The purpose of the assessments, similar to the purpose of risk assessments, is often to facilitate decision making in order to increase capability (Abt et al. 2010, Bier 2001, Johansen & Rausand 2014, Palmqvist et al. 2014).

At the same time, modern societies are becoming increasingly complex (OECD 2003, Calvano & John 2004). Critical infrastructure systems that society depends on, such as the transport system, financial systems and the electricity distribution system, are becoming more interconnected (Rinaldi et al. 2001, Little 2004) and are being transformed into a so-called 'system of systems' that covers vast geographical areas, sometimes crossing national boundaries or continents. Also, the management of these systems is becoming increasingly fragmented (De Bruijne & van Eeten 2007). Since the actors involved may have different objectives and concerns, the management of risk, including assessments of capability, in the context of complexity, uncertainty and ambiguity poses a considerable challenge (IRGC 2005, Bristow et al. 2012).

Increased dependencies and institutional fragmentation mean that the assessment of society's ability to deal with disruptions is becoming practically, as well as methodologically, more difficult. One aspect of this is the need for multi-actor capability assessment. An example of this would be that if the capability of actor A is highly dependent on the performance of actor B, it is important that the capability of actor B is reflected in actor A's assessment. Assume, for example, that actor A is a fire and rescue service and actor B is a hospital. In the case of a serious fire, the capability to save the lives of people trapped in the burning building depends on the performance of both actors. One important consequence of this scenario could be the number of fatalities due to the fire. Clearly, in assessing the number of fatalities, the hospital's capability to treat the people exposed to heat

and smoke is important. However, the hospital's capability to treat victims is highly dependent on the condition of the patients when they arrive at the hospital, which in turn depends on how quickly they can be rescued from the burning building. Thus, there is a need for the two actors to work together and take the dependencies between the two into consideration when assessing capability.

Another difficulty lies in making assessment of capability that are as close as possible to the real capability (Lindbom et al. 2015). Making correct assessments of capability is important since the assessment will guide decisions regarding measures that aim at increasing the real capability. If the estimation is far from reflecting the real capability, the measures might not increase capability at all, instead it might even decrease it.

However, it is difficult to know how well the assessments of capability reflect real capability. There are several reasons for this. The first is that disasters seldom occur and we therefore have little knowledge of the real disaster response capability to compare the assessments of capability to. The second is that the common way of assessing capability makes it difficult to compare the result of the assessment with an outcome of a disaster. Common capability assessment methods are often so called indicator or index methods (Palmqvist et al. 2014). Such a method consists of a list of various aspects (indictors), often in terms of resources available, that should be ticked off from a list. This list of available resources is the result of a capability assessment. If the method is an index method, each indicator is assigned a certain value and a final capability index is calculated, based on some formula, to e.g. 25.4. But, there is no clear connection between such a capability assessment and the outcome of a disaster. I.e. the outcome of a disaster will not be 25.4, but *x* lost lives or a cost of *y* million dollars. Thus, the basis for saying how well capability assessments reflect actual capability by studying capability assessments and comparing them with outcomes from disasters is poor.

As an attempt to rectify this lack of knowledge, this paper presents an experimental study aiming at investigating how well capability assessments reflect real capability in a multi-actor context.

Following this introduction, the outline of the paper is as follows. First we report on previous research on how well capability assessments correspond to the actual outcome. We then present the experiment, including e.g. research questions and hypotheses, participants, tasks and procedure, and analysis method and results. This is followed by a discussion about the findings and a conclusion.

2 PREVIOUS RESEARCH

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 too much energy to think about all choices and all interpretations being made, the human mind uses shortcuts, heuristics, and these affect the outcome when making a decision or assessment (Kahneman 2011). The availability heuristic may cause overestimating the importance of a factor/task/asset due to availability; the anchor heuristic causes overestimation or underestimation due to exposure to numbers; and attribute substitution causes substitution of more complex questions to simplified questions.

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.

Studies with individual capability assessments have shown that people in general underestimate the timeframe of large tasks and overestimate the timeframe for small tasks (Halkjelsvik et al. 2011), underestimate the timeframe to fulfil a task (Dunning et al. 2004), overestimate their driving skills (Mynttinen et al. 2009), overestimate their reading skills (Fredriksson et al. 2011) and a majority describes themselves being above average in ambiguous traits (Dunning et al. 2004). A study which included both individual and peer capability assessments have shown that people of individualistic cultures overestimated their generous manner and underestimated their negative behaviours, although they were about right regarding their peers, while 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). Another study, which included group assessments and group peer assessments, showed that high-achieving groups underestimated their performance while low-achieving groups overestimated their performance, while all groups underestimated all other groups' performances (Sung et al. 2010).

To conclude this chapter, several studies have been carried out on individual and group assessments. However, none of the studies focus on how capability assessments are affected if the actors are dependent on each other in the performance of the task. The experiment described in the next section is an initial attempt to rectify this lack of knowledge.

3 EXPERIMENT

3.1 *Research questions and hypotheses*

The aim of this study is to investigate how well capability assessments reflect real capability in a multi-actor context. Based on this, three research questions have focused the study and a set of hypotheses was created for statistical purposes.

- RQ1: Do the capability assessments match the actual performance?
  - *Hypothesis 1: There is no difference between capability assessments and performances.*

- RQ2: 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 2: There is no difference in accuracy for individual assessments between tasks.*

- RQ3: Is there a difference between how well the capability assessments match the actual performance depending on how the multi-actor dependencies were designed?
  - *Hypothesis 2: There is no difference in accuracy for individual assessments between tasks.*
  - *Hypothesis 3: There is no difference in accuracy for pair assessments between tasks.*

3.2 *Design features*

During the process of developing a suitable experiment, the following features have been crucial for the experimental design:

- *Possible to measure performance.* Capability is in this experiment defined as the ability to perform a specific task and reach a certain result. Therefore, it was important using a straightforward way of measuring performance.
- *Cognitive task.* A cognitive task is considered more valid for future implementation than a time perception task or a physical task. This because a cognitive task more reflects the process of assessing capability in the context of disaster risk management.
- *Dependency models.* In order to design an experiment for multi-actor dependencies, a dependency model was created 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.

In the dependency models 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, or administrative authorities. The actors are considered to have the following characteristics and properties: responsible for different parts of a task, a common goal and limited resources (e.g. knowledge, staff and equipment).

The two actors are dependent on each other in order to reach the common goal. The dependencies are of two categories:

Dependency I: The performance of each actor is essential to reach the common goal, but a poor performance of actor 1 does not affect the performance of actor 2. Example: running relay. If actor 1 underperforms it affects the result, but it does not affect how fast actor 2 is able to run.

Dependency II: The performance of each actor is essential and poor performance of actor 1 affects the performance of actor 2. Example: A scenario where the fire and rescue service and a hospital are working together to save a burn victim (see example in Introduction).

3.3 *Mastermind*

The game Mastermind is played on a game board. To win the game, a code of four dots has 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.

In this experiment, two actors assessed their own and joint capability to solve the code, i.e. at which row they would solve the code at.

3.4 *Participants*

Participants were recruited from Lund University, Luleå Technical University and the senior high school Nils Fredriksson Utbildning, all in Sweden. In total 48 participants participated in the experiment. 27 were engineering students, 15 social science students, 3 interdisciplinary science students, and 3 vocational education students. 26 participants were female and 22 male. Their age ranged from 18 to 48 (median = 22). 60 % of the participants had played Mastermind before, and 30 % were familiar with capability assessment either through military service, sports, university studies or through their occupation. In the experiment the participants were grouped into pairs. Each

participant received a cinema ticket as a thank you for participating.

3.5 *Tasks and procedure*

During the experiment the pairs were randomly assigned to perform one of two tasks. The two tasks represented dependency I (task I) and dependency II (task II) respectively. 14 pairs conducted task I, 6 pairs conducted task II, and 4 pairs conducted both tasks.

In task I each participant had a game board and a code to solve each. The capability of a pair was defined as the sum of both participants' results, see Figure 1. Since the participants in this task solved a code individually, independent individual results were also obtained.

In task II the participants had two joint game boards with one code for each game board. The first participant played the first four rows on game board 1, and then the second participant took over and played the game until the code was solved, see Figure 2. Similarly, the second participant played the first four rows on the game board 2, and after four rows the first participant took over solving that code.

Before performing the assigned task, a ten-minute practice session was held where the participants individually could practice playing Mastermind in order for them to fully understand the game. After the practice session the participants filled in capability assessment forms, assessing at what row they would solve the code. The following forms were filled in, and in the following order:

- Individually filling in the capability assessment form on how they thought they would perform individually (only task I)
- Individually filling in the capability assessment form on how they thought the pair would perform collectively (task I and II).
- In pairs filling in the capability assessment form on how they thought that the pair would perform collectively (task I and II).

Thereafter the assigned tasks were performed. The total time for an experimental session was 45 minutes. The solution of each game board had been randomized prior the experiment. Each experiment leader had six codes to alter between. All codes had four colours.

3.6 *Analysis method*

For the purpose of managing the data from the experiment SPSS was used. The following tests were used to analyse the data: independent samples *t*-test (compares the mean of two independent groups on the same dependent variable), paired sample *t*-test (is used for dependent measurements, e.g. when the same participants are part of both groups that will be compared), Levene's test for equality of variance (tests the hypothesis that the variance in two groups are equal), and Cohen's d for effect size (measures effect size when comparing means, independently from the variables but dependent on which statistical test is used) (Cunningham & Wallraven 2011). The above-mentioned *t*-tests were used because these tests provide an opportunity to determine whether there is a statistically significant difference between the means in two groups.

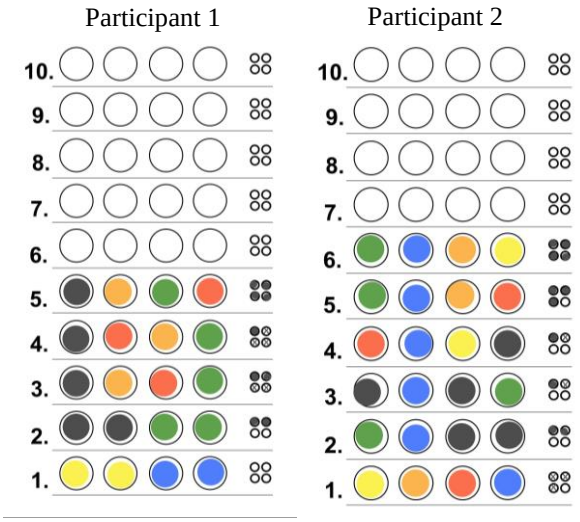

Figure 1. An example result of task I. The individual results are 5 for participant 1 and 6 for participant 2. For task I the result is 11 (5+6).

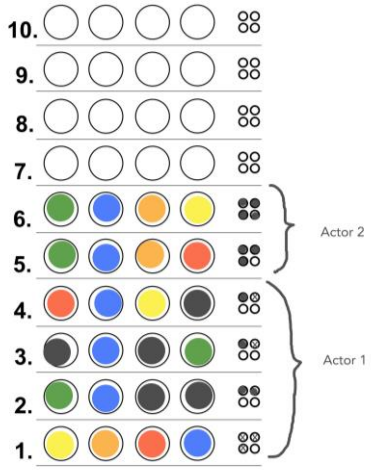

Figure 2. An example result of task II. The result is 6.

3.7 *Results*

- RQ1: Do the capability assessments match the actual performance?

Hypothesis 1: There is <u>no</u> difference between capability assessments and performances.

Five different paired sample *t*-tests were conducted to compare the difference between capability

assessment and performances. The first test showed a significant difference in the scores for the individual capability assessment for the individual performance (M=10.61, SD=3.17) and the individual performance (M=7.11, SD=2.70), $t(35)=4.78$, $p=.000$, $d=.80$. In the second test, there was a significant difference in the scores for the individual capability assessment for the group's performance in task I (M=20.08, SD=6.23) and the group's performance in task I (M=14.22, SD=3.42), $t(35)=4.91$, $p=.000$, $d=0.78$. The third test showed a significant difference in scores for the individual capability assessment for the group's performance in task II (M=10.74, SD=1.52) and the group's performance in task II (M=7.05, SD=1.75), $t(18)=7.32$, $p=.000$, $d=1.69$. In the fourth test, there was a significant difference in the scores for the group's joint capability assessment for the group's performance in task I (M=19.94, SD=5.38) and the group's performance in task I (M=14.22, SD=3.47), $t(17)=3.74$, $p=.002$, $d=0.90$. Finally, the fifth test showed a significant difference in the scores for the group's joint capability assessment for the group's performance in task II (M=10.58, SD=1.61) and the group's performance in task II (M=7.05, SD=1.75), $t(18)=7.17$, $p=.000$, $d=1.65$.

In other words, the capability assessments do not match the actual performance. The participants, both individually, and when assessing in pairs, underestimate their capability for all tasks.

- RQ2: 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 2: There is <u>no</u> difference in accuracy for individual assessments between tasks.

One paired sample *t*-test was conducted that showed a significant difference in the scores for the difference between the individual capability assessment for the individual performance and the individual performance (M=-3.50, SD=4.39) and the difference between the individual capability assessment for the group's performance in task I and the group's performance in task I (M=-5.86, SD=7.17), $t(35)=2.98$, $p=.005$, $d=.59$. An independent sampled *t*-tests showed no significant difference in the scores for the difference between the individual capability assessment for the individual performance and the individual performance (M=-3.50, SD=4.39) and the difference between the individual capability assessment for the group's performance in task II and the group's performance in task II (M=-3.90, SD=2.34), $t(53.95)=.45$, $p=.658$, $d=.12$. Levene's test indicated unequal variances (F=5.01, $p=.029$), so degrees of freedom were adjusted from 54 to 53.95.

In other words, there is a difference between how well the capability assessments match the actual performance depending on if the task is performed individually or in pairs, but depending on the dependency model for the joint performance. An individual makes more accurate assessments of their own capability than of the group's capability for task I. However, the data does not support that an individual makes more accurate assessments of their own capability than of the group's capability when it comes to task II. A qualitative analysis of these results, supporting the quantitative results, will be presented below.

- RQ3: Is there a difference between how well the capability assessments match the actual performance depending on how the multi-actor dependencies were designed?

Hypothesis 2: There is <u>no</u> difference in accuracy for individual assessments between tasks.

The independent sampled *t*-test showed a significant difference in the scores for the difference between the individual capability assessment for the group's performance in task I and the group's performance in task I (M=-7.54, SD=7.04) and the difference between the individual capability assessment for the group's performance in task II and the group's performance in task II (M=-3.90, SD=2.34), $t(34.81)=-2.54$, $p=.016$, $d=-.78$. Levene's test indicated unequal variances (F=12.64, $p=.001$), so degrees of freedom were adjusted from 46 to 34.81.

The individual makes more accurate assessments of the group's capability when it comes to task II compared to task I.

Hypothesis 3: There is <u>no</u> difference in accuracy for pair assessments between tasks.

One independent sampled *t*-tests was conducted that showed no significant difference in the scores for the difference between the joint capability assessment for the group's performance in task I and the group's performance in task I (M=-7.21, SD=6.50) and the difference between the joint capability assessment for the group's performance in task II and the group's performance in task II (M=-3.53, SD=2.144), $t(15.10)=-2.04$, $p=.059$, $d=-.85$. Leve-ne's test indicated unequal variances (F=9.84, $p=.004$), so degrees of freedom were adjusted from 31 to 15.10.

In other words, the individual makes more accurate assessments of the group's capability when it comes to task II compared to task I. For the joint capability assessment, it seems as if there is no difference in accuracy depending on whether the dependencies are of type I or II. However, these results will be discussed below in the qualitative analysis since they showed a significant difference ($p=.027$) when assuming equal variances.

3.7.1 *Qualitative analysis*

A qualitative analysis through a comparison of mean value was made for hypotheses 2 and 3 where the accuracy of individual and group capability assessments was tested. The qualitative analysis of the individual assessments showed that the tasks are ranked, the individual task, task II and task I in accuracy. This is based on that there are significant differences between the individual task and task I as well as between task II and task I and that the accuracy mean value is lower for the individual task than for task II. Hypothesis 3 was not rejected as Levene's test showed that there was not a significant difference between task I and task II. However, a qualitative analysis shows that in capability assessments made in pairs for task II tend to be more accurate than for task I.

4 DISCUSSION

The aim of this study was to investigate how well capability assessments reflect real capability in a multi-actor context 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 the task. 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.

It is our 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 paper.

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 when the participants explained their way of thinking during the experiment. In order to reduce the effects of this phenomenon, the experiment supervisors specifically told the participants that not reaching the estimated row was considered equal to exceeding the estimated row. Despite this encouragement, the participants continued to underestimate their capability.

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) describes as anchoring heuristics.

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.

The main argument to include the individual task as part of task I 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. We found this surprising because it was anticipated that task II 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, we found that there was hardly any time difference between task II and the individual task.

The results of this study suggest that although people find it more difficult to assess capability for task II, than for task I, assessments for task II 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 task 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 results in a more thorough performance by the relevant actors. This suggests that capability assessments with multi-actor dependencies are vulnerable if they are performed with little or no communication.

We believe that the experiment is possible to recreate with similar results. However, due to the nature of Mastermind and the design of the two tasks, it may be difficult to replicate the experiment with other activities than Mastermind. The way task I is designed the performance is often twice the number of rows than for task II. As task I and II 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.

We did not register the results from the practice sessions, but it was apparent that this 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. 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.

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.

The world grows more and more complex with an increasing amount of interdependencies between actors. This in turn increases the risk of misinterpretation of responsibilities and capabilities. Therefore, it will be crucial in the long run to conduct multi-actor capability assessments, in order to meet society's increasing demand for reliable provision of vital services.

In reality, the challenges faced by actors doing capability assessments with multi-actor dependencies are more complex than the experiment conducted in this study. 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 indirectly may result in loss of lives. Therefore, it is critical to assess capability without overestimation or underestimation.

5 CONCLUSION

We have conducted an experiment to study how well capability assessments reflect real capability. A total of 48 participants, randomly assigned into 24 pairs, assessed their individual capability and capability as a group to accomplish a given task. The participants were to assess their capability to perform one of two tasks, and then perform the task. The two tasks differed in how the participants were dependent on each other. In task I, the performance of each actor was essential to reach a common goal, but a poor performance of actor 1 did not affect the performance of actor 2. In task II, the performance of each actor was essential and poor performance of actor 1 affected the performance of actor 2. The participants performing task I also assessed their capability to perform the task individually, and performed the task individually.

Based on the results, we conclude that capability assessments in general underestimate capability, both when performed individually and in group. Further, the results indicate that capability assessments are more accurate for the individually performed task and task II, compared to capability assessment for task I. This indicates that it is easier to predict the results of a task where you are independent or where the results are critically dependent on all actors. Involving all actors may increase the sense of responsibility for all actors to understand the capability and limitation of each actor, including their own, to a greater extent than otherwise. This knowledge is important to take into consideration when developing methods for capability assessment that aims at taking the increasing interdependencies between actors into account.

ACKNOWLEDGEMENTS

We thank the Swedish Civil Contingencies Agency for funding the research on which the present paper is based.

REFERENCES

Abt, E., Rodricks, J. V, Levy, J. I., Zeise, L., & Burke, T. A. 2010. Science and decisions: advancing risk assessment. *Risk Analysis*, 30(7): 1028–1036.

Australian Capital Territory 2012. Emergencies Act 2004, republication no 20.

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*, 95(6): 1252–1267.

Bier, V. 2001. On the state of the art: risk communication to decision-makers. *Reliability Engineering and System Safety*, 71(2): 151–157.

Bristow, M., Fang, L., & Hipel, K. W. 2012. System of systems engineering and risk management of extreme events: concepts and case study. *Risk Analysis*, 32(11): 1935–55.

Cabinet Office 2014. Preparation and planning for emergencies: the national resilience capabilities programme - detailed guidance. Retrieved March 31, 2016, from http://www.cabinetoffice.gov.uk/content/capabilities-programme

Calvano, C. N., & John, P. 2004. Systems engineering in an age of complexity. *Systems Engineering*, 7(1): 25–34.

Cunningsham, D. W., & Wallraven, C. 2011. Experimental design: from user studies to psychophysics [E-book]. Hoboken: CRC Press.

De Bruijne, M., & van Eeten, M. (2007). Systems that should have failed: critical infrastructure protection in an institutionally fragmented environment. *Journal of Contingencies and Crisis Management*, 15(1): 18–29.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(4): 69–106.

Dutch Ministry of Interior and Kingdom Relations 2009. *Working with scenarios, risk assessment and capabilities in the National Safety and Security Strategy of the Netherlands*.

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*, 32(4): 301–321.

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*, 25(2): 314–323.

Homeland Security 2013. *Capability estimation. Comprehensive preparedness guide (CPG) XXX. Pre-decisional working draft. For review purposes only*.

Houdijk, R. 2010. *Regional risk assessment in The Netherlands - an introduction*. The Hague.

IRGC. 2005. *White paper on risk governance - towards an integrative approach*. Geneva: International risk governance council.

Johansen, I. L., & Rausand, M. 2014. Foundations and choice of risk metrics. *Safety Science*, 62: 386–399.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Klein, G. 2009. Conditions for intuitive expertise: a failure to disagree. *American Psychologist*, 64(6): 515–526.

Lindbom, H., Tehler, H., Eriksson, K., & Aven, T. 2015. The capability concept - on how to define and describe capability in relation to risk, vulnerability and resilience. *Reliability Engineering & System Safety*, 135: 45–54.

Little, R. 2004. A socio-technical systems approach to understanding and enhancing the reliability of interdependent infrastructure systems. *International Journal of Emergency Management*, 2(1-2): 98–110.

Ministry of Civil Defence and Emergency Management n.d. National capability assessment. Retrieved March 31, 2016, from http://www.civildefence.govt.nz/cdem-sector/monitoring-and-evaluation/national-capability-assessments/

Mynttinen, S., Sundström, A., Vissers, J., Koivukoski, M., Hakuli, K., & Keskinen, E. 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*, 40(4): 301–309.

OECD. 2003. *Emerging risks in the 21st century: an agenda for action*. Paris: Organisation for economic co-operation and development, OECD.

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 *Proceedings of the Twelfth International Conference on Probabilistic Safety Assessment and Management (PSAM), Honolulu, June 22-27 2014*.

Programme National Security. 2007. *National security, strategy and work programme 2007-2008*. Breda: Programme national security, Ministry of the interior and kingdom relations.

Rinaldi, S. M., Peerenboom, J. P., & Kelly, T. K. 2001. Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Systems*, 21(6): 11–25.

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*, 33(1): 135–145.

Swedish Civil Contingencies Agency. 2014. National risk and capability assessment [Nationell risk- och förmågebedömning]. Retrieved March 31, 2016, from https://www.msb.se/sv/Forebyggande/Krisberedskap/Nationell-risk--och-formagebedomning/

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*, 58(4): 582–592.