PEDAGOGICAE
UNIVERSITATIS
TARTUENSIS
37

TÕNIS MÄNNISTE

Measuring military commanders' decision making skills in a simulated battle leading environment

DISSERTATIONES PEDAGOGICAE UNIVERSITATIS TARTUENSIS

37

DISSERTATIONES PEDAGOGICAE UNIVERSITATIS TARTUENSIS
37

TÕNIS MÄNNISTE

Measuring military commanders' decision making skills in a simulated battle leading environment

UNIVERSITY OF TARTU
Press

Institute of Education, Faculty of Social Sciences, University of Tartu, Estonia

Dissertation is accepted for the commencement of the Degree of Doctor of Philosophy (in Educational Science) on August 28, 2020 by the joint PhD defence committee between the Institute of Education and the Institute of Ecology and Earth Sciences.

Supervisor: Professor Margus Pedaste, PhD
University of Tartu, Estonia

Opponent: Associate Professor / Lieutenant Colonel
Anders McDonald Sookermany, Dr. Philos.
Norwegian Police University College /
Norwegian Defence University College, Norway

Commencement: White Hall of the University of Tartu Museum, 25 Lossi St., Tartu, 9th November 2020, at 1 p.m.

Publication of this dissertation is granted by the University of Tartu

ISSN 1406-1317
ISBN 978-9949-03-458-1 (print)
ISBN 978-9949-03-459-8 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF FIGURES 7
LIST OF TABLES 7
LIST OF ORIGINAL PUBLICATIONS 8
LIST OF ABBREVIATIONS 9
1. INTRODUCTION 10
2. THEORETICAL FRAMEWORK 14
2.1. Military cognitive readiness 14
2.2. The problem and problem solving in the military sphere 16
2.3. Decision-making in the military sphere 18
2.4. Decision-making styles and military decision-making 20
2.5. Expertise and military decision-making 21
2.6. Unit cohesion and military decision-making 24
2.7. Motivation and military decision-making 25
2.8. Measuring military decision-making in the battle context 27
3. METHODOLOGY 30
3.1. Phase 1: Preparation for instrument development 30
3.2. Phase 2: Instrument development, data collection, instrument quality check and improvement based on empirical data 34
3.3. Phase 3: Describing the results of the SJTDM based on different groups 45
3.4. Phase 4: Finding factors predicting results of SJTDM 45
4. RESULTS 48
4.1. Phase 1: Preparation for instrument development 48
4.1.1. Step 1. Specific characteristics of a problem and problem solving in military sphere 48
4.1.2. Step 2. Types of instruments/tests, selecting the instrument for current doctoral thesis 49
4.2. Phase 2: Instrument development, data collection, instrument quality control and improvement based on empirical data 50
4.2.1. Instrument development and data collection 50
4.2.2. Instrument quality control and improvement 50
4.3. Phase 3: Describing the results of the test based on different groups 53
4.4. Phase 4: Factors predicting results of SJT 55
5. DISCUSSION 60
5.1. Knowledge needed to develop a test for assessing decision-making skills in military context 60
5.2. Quality of the compiled test for measuring military commanders' decision-making skills 63

5.3. Decision-making skills of the military commanders 64
5.4. Predictors of military commanders' decision-making skills 65

6. CONCLUSIONS 69
6.1. Summary of main conclusions 69
6.2. Limitations of the doctoral thesis 70
6.3. Implications 72
6.3.1. Theoretical implications 72
6.3.2. Practical implications 72

7. REFERENCES 74

8. ANNEXES 82

9. SUMMARY IN ESTONIAN 89

10. ACKNOWLEDGEMENTS 94

11. PUBLICATIONS 95

CURRICULUM VITAE 251

ELULOOKIRJELDUS 253

LIST OF FIGURES

Figure 1. Research design. Page 31
Figure 2. Flowchart of the literature search procedure. Page 33
Figure 3. Conceptual model for predicting platoon level military commanders' performance on the SJTDM (Article **V**). Page 47.
Figure 4. Conceptual model with standardized path coefficients (decision-making skills is a latent variable, all others are aggregated variables) (Article **V**). Page 59.

LIST OF TABLES

Table 1. The schema by which points were given in the case of matching tests. Page 39.
Table 2. Descriptive data of the sample (Article **III**). Page 41.
Table 3. The internal consistency of SJTDMs before and after modfications based on the IRT analysis. Page 50.
Table 4. 1PL IRT analysis results to describe 24 test items (n = 214) (Article **III**). Page 52.
Table 5. Results of independent samples analyses (table compiled based on the results presented in Article **IV**). Page 54.
Table 6. Results of dependent samples analyses (table compiled based on the results of Article **IV**). Page 55.
Table 7. Goodness of fit data of the measures used in the doctoral thesis (based on CFA models) (Article **V**). Page 56.
Table 8. Fit indices of the Structural Equation Models (Article **V**). Page 56.
Table 9. Estimates of conceptual model (standardized coefficients) (Article **V**). Page 58.

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, which will be referred to in the text by their Roman numerals:

I. Männiste, T. & Pedaste, M. (2015). Probleemi mõiste sõjanduses ning probleemõppe lõimimisvõimalustest sõjaväelise väljaõppega. *KVÜÕA toimetised, 20*, 198–233.

II. Männiste, T., Pedaste, M. & Schimanski, R. (2019). Review of Instruments Measuring Decision Making Performance in Military Tactical Level Battle Situation Context. *Military Psychology*, *31*(5), 397–411. https://doi.org/10.1080/08995605.2019.1645538

III. Männiste, T., Pedaste, M. & Schimanski, R. (2019). Situational Judgement Test for Measuring Military Tactical Decision Making Skills. *Military Psychology*, *31*(6), 462–473. https://doi.org/10.1080/08995605.2019.1664366

IV. Männiste, M., Rajaste, R., Suviste, R. & Pedaste, M. (2018). Olukorrapõhised otsustustestid rühmataseme sõjaväeliste juhtide otsustusvõime mõõtmiseks lahingujuhtimist matkivates olukordades. *Sõjateadlane, 9*, 63–118.

V. Männiste, T., Pedaste, M., Kasearu, K. & Schimanski, R. (2019). Predicting Decision-Making Skills among Estonian Platoon-Level Military Commanders. *Military Psychology*, in the process of review. Submitted 15.08.2020.

The author of the present dissertation contributed to these publications as follows: **I**– **V**: developed the concept, planned and carried out data collection and testing, participated in data analyses, and wrote the articles as primary author.

LIST OF ABBREVIATIONS

CFA – Confirmatory Factor Analysis
EDF – Estonian Defence Forces
EKV – Eesti Kaitsevägi (Estonian Defence Forces)
ENDCOL – Estonian National Defence College
GDMS – General Decision Making Style
IRT – Item Response Theory
MAWS –Motivation at Work Scale
MDMP – Military Decision Making Procedure
SEM – Structural Equational Modelling
SJT – Situational Judgement Test
SJTDM – Situational Jurdgement Tests Measuring Decision-Making Skills

1. INTRODUCTION

During the period after the Cold War (between 1990–2008) many European countries lost the ability to defend themselves in the event of a war (Hedlund, 2019). However, after the events in Georgia in 2008 and Ukraine in 2014, a number of NATO members, and particularly those on NATO's eastern border, are reintroducing conscription and looking for ways to improve the performance of their reserve soldiers (Mohdin, 2018).

One such small country is the Republic of Estonia, where every year about 3500 new conscripts begin compulsory service in the Estonian Defence Forces (EDF) to become reservists and be ready to defend the country if needed. During their service, they need to learn several skills, e.g. practical skills but also higher-order cognitive skills as decision-making skills. The EDF are modest in size and can thus only rely on excellent tactics during combat situations. This, however, means that all the soldiers, and especially commanders with the best decision-making skills, are crucial for the EDF to stand a chance against a possible enemy attack. Shortcomings in soldiers' and especially commanders' education cannot be compensated for by the size of the army. In the EDF, about 1000 conscripts of the overall 3500 begin their compulsory conscription three months earlier than the others and receive special training to become reserve commanders of small (up to 30 men) military unit (Kaju, 2013). In command of that military unit (squad or platoon), reserve commanders are physically closest to the enemy in battle and carry out the fighting. In addition, cadets of EDF and therefore future professional officers are selected and developed at the conscription level, with pre-call being the primary source of cadet candidates – thus, conscription and conscripts are very important for EDF.

Leading their unit in battle is the main task of the squad, platoon and even company-level military commanders. It is of course an utmost difficult and demanding task for every military officer. In other words, they have to learn how to solve problems in the battlefield with using deadly force against the enemy. That is because it demands courage and suitable attitude to act independently, presumes the need to think critically and take initiative when needed, while at the same time still fulfilling the task given by the higher commander. This appears to be the educational ideal, which forms the basis for the acquisition of skills in modern professional military communities of practice (Sookermany, 2012). To be successful in battle, skilful commanders need to form a plan of action and make adjustments as events unfold, with the end goal of maximizing exploitation of the advantages over the enemy, or minimizing harm done by the enemy. Lussier and Shadrick (2004) have pointed out that while performing their duties, small unit military commanders need to assess the situation, scan for new information, deal with individuals under stress, and monitor the progress of multiple activities of a complex plan. This means that commanders are under constant pressure to multitask and deal with simultaneous problems of different types and magnitudes. It is very demanding even for professional military commanders, not to mention inexperienced conscripts. Therefore, the commanders' education is very important.

However, given the modest size of EDF and the need for skilful reserve commanders mentioned earlier, combat performance of conscripts who become reserve commanders by the end of conscription must be at least close to that of professional military commanders. It poses the main challenge for the current doctoral thesis as well – in the Estonian Defence Forces, there is a need to teach pre-call conscripts in a limited time to be able to take on the role of commanders. Therefore, we need to understand what are the possibilities to design an effective educational path for conscripts' education in a situation where there is only 11 months available for training early-call conscripts. During this training, conscripts have to be prepared to become wartime commanders of small military units of up to 30 men (squad or platoon). This necessitates that the EDF provide conscripts with the best possible training during conscription. Officer education has traditionally been focused on learning to solve current problems, situations, and tasks; while academic education focuses more on learning for the future and for problems, situations, and tasks, we know little or nothing about yet (Hedlund, 2019). Conscripts are not in the position to get academic education during mandatory service time, and the EDF have therefore recently come to place great value on human resources and, in recent years, started a long-term project aimed at studying and developing human resources (Allik & Talves, 2016). Similar views about the US military have also been proposed by some researchers in the US (Dees, Nestler & Kewley, 2013).

The human resource project, which was carried out in Estonia since 2015, concentrates on mapping different aspects of conscripts' profile. The tool for data gathering includes many categories (social, health, motivation etc.) that can be later used in different predictive models (Allik & Talves, 2016). Contributing to this project by using its data for predictive purposes in the context of testing training outcomes was also considered one purpose of this doctoral thesis. However, this project did not focus on evaluating conscripts' decision-making skills that are definitely very important in their education. Even more, it turned out that there is a lack of appropriate instruments for measuring their decision-making skills in the military context. Without this instrument, it would be not possible to design studies to test the effect of different interventions on conscripts' decision-making skills.

One way to quickly improve the proficiency of the commanders is through acting/practising in a real workplace under real conditions as an apprentice (Sookermany, 2012). It is complicated, though, because in the military profession, it is impossible during peacetime for commanders to experience decision-making in actual battle situations while leading the unit. Professional soldiers can be sent on missions abroad in order to gain battle experience, but this is not an option for conscripts chosen to become future reserve commanders. Additionally, battle experience during missions might not be enough, because such missions take place in quite different types of terrains and environments. As a result, by the end of conscription, reserve commanders have not experienced battle other than in a few exercises and thus might not have a clear picture of what leading troops in battle really means.

This seems not to be the case with only inexperienced young conscripts. Even professional soldiers have problems understanding what constitutes a good military skill utilisation (Sookermany, 2012). A lack of opportunities to participate in combat operations will deprive the soldiers of the opportunity to try out their own skills in real situations and over time build up the experience required to carry out tasks they are given. This shows that officers' education is a challenge not only in Estonia, but in other NATO countries as well.

Nevertheless, given the small size of the EDF, in case of military conflict, it can only rely on excellent performance – commanders ultimately performing better than the enemy. Because of this, their training must be of excellent quality, and has to ensure the ability to make successful decisions in critical (battle) situations.

Without being able to reliably measure or thus know the level of training outcomes of commanders, it is not possible to reasonably conduct or improve commanders' training. Currently, there is no measure for assessing the quality of the decision-making skills of platoon level commanders. Thus there is nothing that the improvement of training can be placed upon. Because of that, a need exists for a suitable instrument for measuring platoon leaders' decision-making skills in critical battle-like situations. Due to the time limit, most of the conscription time should be allocated to training purposes. Thus, the evaluation itself should not be resource- and time-consuming, expensive, labour-intensive and difficult to prepare. Nevertheless, it should still yield valid results.

Considering the above, the goal of this PhD project was to construct an instrument for measuring decision-making skills in battle-like situations and test its suitability based on a sample consisting of EDF conscripts and cadets, and thereafter determine the predictors of a better test result. In order to achieve the goal, five research questions were raised and accordingly, five studies were carried out in order to answer those questions.

Firstly, interest was shown toward the type of problems that low-level military commanders might face in battle situation. It was done by comparing the differences between the concept and the characteristics of the problem in the military context to commonly used concepts and characteristics. Thus, the first research question was formulated as follows: (i) **What are the specific characteristics of the problem and problem-solving in the military profession and how does it deviate from the general definition of a problem?**

Secondly, before it is possible to improve the training, it is necessary to know the current level of conscripts' decision-making skills. For this, a reliable instrument is needed, enabling to measure decision-making skills of conscripts. As EDF is very small and has very limited resources available, it would be reasonable to reuse an already existing instrument. If there is no suitable instrument available, the compilation of a new instrument should not be expensive, labour-intensive, or difficult. In order to save most of the valuable conscription time for training purposes, the testing should not be resource- and time-consuming. It is not known whether such an instrument measuring decision-making skills in military tactical-level battle context exists at the moment. Thus, it seems reasonable firstly to find out, how decision-making skills in critical situations have been measured, and what specific instruments have been used to measure it. Therefore, the second research question was formulated as follows: (ii) **What instruments have been used for measuring decision-making skills in battle (critical) situations and which instruments are best applicable and scientifically credible for measuring military commanders' decision-making skills?**

After the existence (or non-existence) of a suitable instrument is confirmed, it seems reasonable to start adapting (if exists) or creating (if does not exist) an instrument to measure decision-making skills of low level military commanders. This is needed to develop, test and validate an instrument which enables measuring a platoon level military commander's decision-making skills in battle situation. Therefore, the third research question of the thesis was as follows: (iii) **How to adapt the instrument for EDF purposes and what is its quality?** Thus, the next aim of the study was to construct an instrument for measuring platoon-level military commanders' decision-making skills in battle-like situations and assess the quality of the constructed test.

After constructing, quality checking, validating and thus improving the instrument, it became possible to assess the differences in decision-making skills between more advanced (cadets) and novice (conscripts) test-takers. Therefore, the fourth research question of the thesis was as follows: (iv) **How accurately does the instrument distinguish decision-making skills of novices from experts/professionals in a simulated platoon leader battle scenario in the example of EDF cadets (advanced) and conscripts (novices)?** Thus, this sub-study aimed to find out how good the developed instrument is in distinguishing between novice and advanced military personnel and how the testing format influences the outcomes.

Finally, in connection with the EDF human resource project (Allik & Talves, 2016), it was possible to study the test performance predictors. Therefore, the fifth research question was as follows: (v) **What are the attributes that predict a better result on the decision-making test based on the sample of EDF cadets and conscripts?** Thus, the final sub-study aimed to identify the predictors of better performance in decision-making test in a simulated battle-leading environment.

2. THEORETICAL FRAMEWORK

This chapter gives an overview of basic concepts that underline current doctoral thesis. It provides most important definitions that begin with the concept of military cognitive readiness, which underlines the elements of soldier performance. Based on that, problem solving and decision-making as important elements of military sphere are described followed by other important aspects of military decision-making.

2.1. Military cognitive readiness

The theoretical construct of military cognitive readiness is a scientifically sound approach that has been used to describe the performance of military personnel for decades. It was selected as the main theory for the current doctoral thesis because it has been seen as having three important elements: (i) it is a predictor of cognitive performance; (ii) requires tolerance of an uncertain, demanding and stressful environment; (iii) is an interaction between the individual and the anticipated situation (Morrison & Fletcher, 2002).

Military cognitive readiness has been defined in many different ways, but the oldest definition is offered by Morrison and Fletcher (2002), which emphasizes on requirements to perform effectively in a military operation. They defined military cognitive readiness as mental preparation (including skills, knowledge, abilities, motivations, and personal dispositions) and individual needs to establish and sustain competent performance in the complex and unpredictable environment of modern military operations (Morrison & Fletcher, 2002).

Due to the large variety in of military sphere, Grier (2012) has recommended a classification of military cognitive readiness in her review article, which in addition to individual's preparedness (mostly stable in time) also includes elements of individual state, which are more changing in time: strategic, operational and tactical. By its nature, it thus follows the logic of levels of war (Macgregor, 1992).

Based on Grier (2012), the definitions of military cognitive readiness are as follows:

- Tactical level cognitive readiness: a state of mental acuity for ensuring an acceptable level of performance during assigned missions. It emphasizes the state of an individual (e.g., stress, workload, and motivation) during an ongoing military operation.
- Operational level cognitive readiness: definition is the same offered by Morrison and Fletcher (2002) above. Operational cognitive readiness consists of a combination of attributes (i.e., knowledge, skills, and abilities) and states (i.e., motivations) to ensure an individual can perform optimally while deployed.

- Strategic level cognitive readiness: an individual's potential to perform assigned cognitive task in the complex and unpredictable environment of modern military operations. It emphasizes the competencies and aptitudes of the individual (e.g., cognitive capabilities, personality).

As proposed by Macgregor (1992), the levels of war are merging in the context of modern war, so the definitions of military cognitive readiness should be also merging. In this doctoral thesis, the definition of original military cognitive readiness (Morrison & Fletcher, 2002) is preferred; however, a short overview of operational cognitive readiness is also given, because the definitions used are exactly the same.

According to the original approach, military cognitive readiness consists of ten psychological components: situation awareness, memory, transfer of training, metacognition, automaticity, problem solving, decision-making, mental flexibility and creativity, leadership and emotion (Morrison & Fletcher, 2002).

Problem solving in the context of military cognitive readiness is viewed as an activity that transforms goals and sub goals into a plan of action by processes such as trial-and-error, proximity, fractionation and knowledge-based referrals. Decision-making in the context of military cognitive readiness is described as the selection of tactical and strategic plans, which are frequently primed by the recognition of learned patterns (Morrison & Fletcher, 2002).

However, Grier (2012) in her review of literature offers an even more detailed set of constructs that fit under the umbrella of military cognitive readiness:

1. Knowledge that is further divided into three categories: military knowledge, deployment knowledge, and general knowledge.
2. Cognitive functions that is further divided into eight categories: decision-making, problem solving, planning, analysis, judgment, systems perspective (i.e., awareness of indirect effects), critical thinking, and metacognition.
3. Expertise, which is further divided into four categories: situation awareness, pattern recognition, sense making and automaticity.
4. Motivation, which is further divided into three categories: general, locus of control, self-efficacy.

Problem solving and decision-making seem to be the most relevant constructs for further consideration in the context of the current doctoral thesis. However, in Grier's model of military cognitive readiness, both problem solving and decision-making are placed under the umbrella of cognitive functions. Grier (2012) herself considers decision-making and metacognition as the most important categories, because they inherently include the other concepts (Grier, 2012). This claim seems strange, because decision-making logically seems to be a part of problem solving. However, there are many different concepts and sometimes problem solving and decision-making are interrelated. Greir (2012) seems to follow the logic, where problem solving and decision-making differ by the nature and time pressure of the situation as proposed by Zeichmeister and Johnson (1992). In this case, problem solving and decision-making are treated almost as synonyms. By this concept, problem solving is a situation where selecting solutions is done in a stable environment which allows for slow and deliberate processing (Zeichmeister & Johnson, 1992). Decision-making, according to this approach, means almost the same, e.g. selecting courses of actions, but it happens in a highly complex and dynamic environment (Zeichmeister & Johnson, 1992).

However, in contrast to Zeichmeister and Johnson (1992), some authors define problem solving much more widely than just a selection of solutions in a stable environment. For example, Anderson (2015) has defined problem solving as "goal-directed behaviour that often involves setting sub goals to enable the application of operators" (Anderson, 2015, p 183). By operators, Anderson (2015) means intermediate actions that transform one state of the problem solving into another (sequential) state.

As is obvious, those two definitions differ greatly and as they have been developed for normal civilian life, they do not even take into account conditions specific to military sphere. Thus, in the following sections, problem solving and decision-making are looked at from the military perspective.

2.2. The problem and problem solving in the military sphere

One important component of military cognitive readiness is problem solving. Based on Grier (2012), classification of military cognitive readiness, it belongs to the category of cognitive functions. Article **I** of the current doctoral thesis concentrated on the definitions and characteristics of the problem and problem solving in military sphere. However, main definitions of the problem and problem solving are provided here as well.

It is argued that it is possible to even discuss the definition of a problem only in the context of human relationships, as proposed by Meacham and Emont (1989). Without a human being as the one who has to solve the problem, there cannot be any problems. In addition, problem has mostly been defined (up to some extent at least) in the context of problem solving and quite often it has been defined quite generally, describing only general steps of problem solving without taking into consideration the precise context (Meacham & Emont, 1989).

Thus, problem solving serves the need of every human being (living system) to adapt and survive while connected with other closed and living systems (Sinnott, 1989). As such, there are many definitions of a problem. Some of these argue that a problem only arises in the case of some need and thus they define problem as someone's need to achieve some goal without knowing how to do it (Mayer & Wittrock, 1996). However, in most cases a problem has been defined as a gap between the current state and the desired state that has to be eliminated by the solver (Jonassen, 2000). Jonassen adds that it is important that problem solver is able to recognize the current and desired end states, otherwise there cannot be a problem for this particular person (Jonassen, 2000).

Robertson (2001) adds some important elements to this definition and argues that a problem only exists when the solver does not know immediately how to eliminate the problem and while doing so, he/she needs to act (Robertson, 2001). In addition to just acting, Eysenc (1984) has argued that it also presupposes the need to think and use at least some cognitive processes (Eysenck, 1984).

While continuing to look at the definition of the problem from the solver's perspective, Jonassen (2000) has argued that he/she needs to be motivated to solve the problem and it can only happen, if solving that particular problem offers cultural, social or intellectual value to the problem solver (ibid). Other researchers (Arlin, 1989; Nitko, 2001) have also highlighted the importance of the value of the problem to the solver, emphasizing that only this motivates the solver to act.

Davis (Davis, 1973) has emphasized some circumstances where a problem does not exist and argues that it is a case where the solver has no motivation to solve the problem, when it is possible for the solver to ignore the problem, when the solution is too obvious to the solver or it is possible to solve the problem by using the trial-and-error method only. Robertson (2001) adds that there cannot be a problem when the solution is known to the solver in advance (Robertson, 2001).

To conclude, a problem occurs when:

- There is a recognizable gap between the current and goal state in some situation.
- There is a need to eliminate this gap and it has to serve some purpose, the solution and the process of eliminating the gap have to be unknown.
- There is a need to act in order to eliminate the gap and achieve the goal, it has to include thinking and other cognitive processes, and the use of creativity is also encouraged.
- Eliminating the gap and achieving the goal offers some cultural, social or intellectual (or those combined) value to the solver.

A problem does not occur when:

- The solver does not recognize the gap between the current and goal state.
- It is possible for solver to ignore the gap and he/she decides to do so.
- Eliminating the gap does not offer any cultural, social or intellectual value to the solver.
- The process of eliminating the gap and/or desired goal state are familiar to the solver.
- Eliminating the gap is possible without thinking by using simple mechanical actions or just the trial-and-error method.

In the case of the military, those definitions seem to lack something, because it is obvious for example that in case of an enemy attack the problem will not go away if we just stop caring about it. The other issue in the military is that in case of a wartime problem situation, there is almost never only one solver who has to deal with it, but units of different sizes are tasked with solving problems. Within these groups, a strict chain of command exists, and it must not be violated. This means that problems must be solved regardless of whether the individual sees them as a problem or not and it is not as important if it offers any cultural, social on intellectual value.

It is also important to point out that even if the problem solver finds a way to ignore the problems or just does not recognize them in a war, they might escalate and cause unnecessary suffering for others, e.g. failing to identify mines in mine fields, will most probably lead to human casualties. This aspect in military problem-solving is especially important, as it is unique to the military field. In case of military conflicts, lethal weapons are used to achieve one's goals, and this must be taken into account when discussing problem-solving in the military. Therefore, in this doctoral thesis, but especially in Article **I**, some modifications are proposed in the definition of a problem and problem solving in order to adjust it better to the military context.

2.3. Decision-making in the military sphere

Another important component of military cognitive readiness is decision-making. The research in decision-making focuses on explaining preferential choice and actions (Connolly, Arkes & Hammond, 2000; Hastie, 2001) and embodies different approaches for conceptualizing and measuring it (Jonassen, 2012).

Based on Grier's (2012) classification, decision-making belongs to the category of cognitive functions together with problem solving. This classification is backed up by concept, which many researchers possess. In other words, they see problem solving and decision-making as synonyms with the exception that one happens in a stable environment and other in a highly complex environment (Zeichmeister & Johnson, 1992). However, there are other viewpoints available, which consider decision-making as a critical component within complex problems (Jonassen, 2012).

Such complex problems are believed to be, for example, negotiation, design, diagnosis and command and control (Means, Salas, Crandall & Jacobs, 1993). Among the ones mentioned, command and control falls into the military sphere. Thus decision-making in the military, especially in the case of war, should be treated as an important part of complex problem solving, which may occur repeatedly while solving a particular problem. Yet another distinction has to be made in the case of wartime problem solving. In other words, analytical and more intuitive types of decision-making concepts and activities have to be clearly distinguished.

This claim is supported by the fact that research on decision-making has historically been divided into two distinct conceptions: normative or prescriptive (e.g. analytical) and descriptive or naturalistic (e.g. intuitive) models (Jonassen, 2012). Normative models of decision-making theories are based on the assumptions that decision-makers are rational individuals who are focused on identifying and acting on the optimal choice in a set of possibilities for every situation (Jonassen, 2012). On the other hand, naturalistic decision-making theories argue that people rarely act in such a rational way, and that they are instead motivated and influenced to a great extent by subconscious drives and emotions and rely on personal identities and social expectations in their decision-making (ibid).

This historical distinction between analytical and naturalistic decision-making also applies in the military command and control context (Vowell, 2004). In the military context, a unique term for decision-making procedure is used, which is called the military decision-making procedure (MDMP). For example, in the US, this process is prescribed in the Army Field Manual 5–0. Similar processes of MDMP are used in all NATO countries (typically with some small modifications) and also in many countries that are currently not members of NATO. MDMP is a planning tool that establishes techniques for analyzing a mission, developing, analyzing, and comparing courses of action against criteria of success and each other, selecting the optimum course of action, and producing a plan or order (US.Army, 2010). It is a highly complex process in itself that incorporates multiple staff estimates, continuous intelligence inputs; and the outcome is usually a very detailed operations order (Vowell, 2004).

It follows that MDMP is exclusively a tool for military headquarters (battalion and higher). Regardless of the command level, MDMP can be seen as just the tools developing a number of courses of actions that are then compared with the best option being chosen to achieve a higher commander's goal (Saaty, 2008). As such, MDMP follows the logic of analytical decision-making and although it has been criticized for being too time consuming, it remains important and it should be always considered as good option for planning the battle, when there is enough time available (Saini, 2008).

However, after plans have been made and orders issued, the actual fighting takes place far away from the headquarters and is typically conducted by low level (company and lower) military commanders. In this low level, small unit commanders need to quickly form out of given order a plan of action, then make adjustments as events unfold, with the end goal of maximizing exploitation of the advantages over the enemy or minimizing harm done by the enemy. While doing that, military commanders constantly need to assess the situation, scan for new information, deal with individuals under stress, and monitor the progress of multiple activities of a complex plan (Lussier & Shadrick, 2004). As such, it is often ad hoc decision-making, relying primarily on experience and intuition, because long planning procedures are not even possible. In this case the naturalistic decision-making models have to be used. The most familiar model that has been used in the military context is the recognition primed decision-making model, which was developed as a result of Gary Klein's (2000) work in naturalistic decision-making (Klein & Klinger, 2000). Naturalistic decision-making has also some limitations. It has been argued that naturalistic decision-making requires expertise and a sufficient knowledge base, so it is not recommended if the decision-maker has little experience in a situation (Vowell, 2004). That is because in the case of ad hoc decisions, future events are predicted and countermeasures chosen on the basis of information at hand and experience gained in the past (*ibid*).

It follows that in the military sphere, a clear distinction can be made between the analytical decision-making process (e.g. making plan for battle) and, the naturalistic decision-making process (e.g. carrying out the plan in the real battle). It ought to be emphasized, though, that analytical and intuitive approaches are not mutually exclusive types of decision-making, but rather they stand in complement (Sadler-Smith & Shefy, 2004). Focusing on intuitions does not mean that conscious analytical thinking is ignored. Intuition and analysis thus work in tandem (Klein & Klinger, 1991) and both are equally important in making good competent decisions.

Nevertheless, since the focus of this doctoral thesis is on the lower military command level – e.g. on small unit commanders' decision-making during battle – naturalistic decision-making is more important. Naturalistic decision-making at the low tactical level, while leading a unit in battle, is an individual action that is related to a number of individual skills and traits. In addition to level of expertise (earlier experience and adequate knowledge), a commander's actions are influenced by decision-making styles, situation awareness, intuition, self-efficacy, and motivation, among other qualities.

Naturalistic decision-making and intuitive decision-making are often considered synonymous. The closest scientific instruments that measure intuition in the context of decision-making, are decision-making styles. The next section gives a short overview of decision-making styles.

2.4. Decision-making styles and military decision-making

As mentioned earlier, naturalistic (e.g. intuitive) decision-making in the context of low level military commanders battle leading is the interest of the current doctoral thesis. The closest scientific instrument, which deals with decision-making styles, including intuitive decision-making, is the measure of decision-making style. According to Rowe and Mason (1987), decision-making style refers to the way a person uses information and derives meaning from it in the process of decision formulation (Rowe & Mason, 1987). For example, previous research on fire-fighters has shown that more skilled personnel differ in their decision-making style to their less skilled counterparts (Cohen-Hatton & Honey, 2015).

There is ongoing debate about whether decision-making styles are stable (e.g. closer to the personality traits or cognitive styles) or situation specific and do they change over long period of time (Berisha, Pula & Krasnigi, 2018). However, many researchers have argued that even if individual decision-making style differs a little by situation due to individual characteristics, people still have one dominant style (Scott & Bruce, 1995; Thunholm, 2004).

A common measure of decision-making style that has often been used in the military is the General Decision Making Style (GDMS) inventory (Scott & Bruce, 1995). According to this inventory, there are five different decision-making styles: *rational style* characterised by thoughtful consideration of alternatives and evaluation; *intuitive style* characterised by decision-making based on inner feelings and premonition; *spontaneous style* characterised by feelings of immediacy and need to make decision quickly; *dependent style* characterised by seeking lots of help and advice from others before making a decision; and *avoidant style* characterised by procrastination and a strong disposition to avoid decision-making altogether. Thunholm (2009) has tested GDMS inventory on military personnel to see if any systematic differences in decision-making styles exist between military team leaders and team members. The results showed that team leaders in the military are higher in spontaneous and intuitive decision-making style, and lower in dependent, avoidant and rational decision-making styles. Some other studies have also argued for the importance of intuition in leaders' and executives' decision-making (Sadler-Smith & Shefy, 2004). Though Thunolm's study was based on self-reporting, it also argued that self-reported GDMS results do not differ significantly from observed GDMS results (Thunholm, 2009), meaning that GDMS inventory is a reliable source of information for studying decision-making styles.

However, the theoretical foundations between different decision-making styles and inventories are yet unclear. Rational style can be identified as preference for analytical decision-making style, and intuitive decision-making style can be identified as behavioural decision-making style in other popular Decision Style Inventory (DSI) (Rowe & Mason, 1987). However, Berisha et al (2018) compared those two instruments and found that there was no convergent validity between them.

However, decision-making styles alone seem not to be enough for making good decisions in stressful situations. Something else is needed, which based on Grier's (2012) classification of military cognitive readiness, is expertise. The next section gives a short overview of expertise in the context of the military.

2.5. Expertise and military decision-making

Expertise is one of the categories of military cognitive readiness and regardless of the domain, decision-making skills of an individual are dependent on the expertise a person has (Grier, 2012). Rasmussen (1983) has defined expertise as the ability to assess situations and choose the optimal action quickly using few cognitive resources. Being an expert means that much of the knowledge, which is gathered through effortful practice, resides within the unconscious and surfaces only when the individual takes an action or makes a decision based on "feel" or "intuition" (Bennet, Bennet & Avedisian, 2015).

There are many studies that bring out the effect of expertise on decision-making: for example in the case of successful executives in civilian organisations (Sadler-Smith & Shefy, 2004), as well as in the case of military (Vowell, 2004). It is also known that expertise is mediated by, and developed through experience, and thus practice is required to become an expert. Simply having knowledge of the field is not sufficient to become an expert, as a person can be knowledgeable on a topic, but not perform at expert levels (Grier, 2012; Norman, 2006). Skilled or expert commanders who have greater situational awareness tend to rely more on their intuitive understanding of the situation and make instant decisions based on their "gut feeling" and previous experience of similar situations. In other words, more skilled commanders tend to rely more on intuitive decision-making. Novice commanders, on the other hand, tend to assess and analyse the present situation more carefully and consciously before making decisions (Cohen-Hatton & Honey, 2015). It seems to be in line with the notion that naturalistic models of decision-making are the most suitable for those who have gained expertise in the field (Vowell, 2004).

In describing the process of becoming an expert, Rasmussen (Rasmussen, 1983) has identified expertise as a three-step process: (i) the knowledge based level, where decision-making requires a thoughtful analysis and thorough consideration; (ii) rules based, whereby the individual makes use of (known to him/her) rules to deduce the decision; and (iii) skills-based level, whereby the process of decision-making is automated and an individual makes use of previous patterns and tacit knowledge. For example, studies in chess, physics, and problem-solving show that whereas novices are able to comprehend superficial task-related problems, experts are able to analyse problems in a more in-depth manner, categorise them and offer more complex solutions (Fuglseth & Grønhaug, 1995; Glaser, 1985).

The levels of becoming an expert (Rasmussen, 1983) seem to be suitable in the case of decision-making, whereas consciously thinking and analysing the situation for making optimal decisions is imperative for novices (Sadler-Smith & Shefy, 2004), but doing that can actually degrade an expert's decision quality (Klein & Klinger, 1991). This idea has been the objective of the work done by Cohen-Hatton and Honey (2015), whose study found that conscious thinking before decision-making did not affect expert firefighters' reaction time as it did in the case of beginners.

Tactical battlefield problems also tend to be viewed differently by military experts and novices. Novices often regard them as puzzles, which have "school book" solutions, while more experienced officers view them in a wider context, acknowledging the possibility that the enemy may not always react as expected to a predictable course of action (Cohen & Thompson, 1999). It follows that expertise should be a good predictor of better performance. Grier (2012) has identified many topics, which are relevant to military expertise with regard to military cognitive readiness: situation awareness, pattern recognition, sense making and automaticity.

A concept closely related to decision-making is situation awareness, which was treated as a component in the original concept of military cognitive readiness (Morrison & Fletcher, 2002). However, Grier (2012) did not include situation awareness as a separate component in to her concept of military cognitive readiness due to the fact that there is a debate whether situation awareness is a process or a product (Salmon et al, 2008) and as such, situation awareness is not a skill but rather something, that whether exist or does not exist in a moment of time (Grier, 2012).

However, as situation awareness is a component of utmost importance for decision-making in the context of expertise and it helps to integrate many other important components of military cognitive readiness and expertise. Situation awareness is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1997, p. 17).

As already mentioned, there are two conflicting ways in which situation awareness can be understood (Saner, Bolstad, Gonzalez & Cuevas, 2009). For some researchers, situation awareness means the process of acquiring information about one's surroundings (Gorman, Cooke & Winner, 2006). Another way to understand situation awareness is that it is a mental state of being knowledgeable about the surrounding environment to a certain degree (Endsley, 1995). Authors who have the latter type of situation awareness in mind typically make a distinction between situation awareness as the *state* of being knowledgeable and situation assessment as the process of attaining that knowledge about the surrounding environment.

For the purposes of the current doctoral thesis, the latter way of understanding situation awareness is preferred. The most famous model of situation awareness is Endsley's situation awareness model, based on which situation awareness can be understood as a three-level process (Endsley, 1995):

(i) Level 1: perception of the elements in the environment;
(ii) Level 2: comprehension of the current situation;
(iii) Level 3: understanding the dynamics of the situation and (accurate) estimation of future developments of the situation.

Situation assessment in this case can be understood as the process of acquiring situation awareness and situation awareness, in turn, as the state of being familiar with the surrounding environment.

In the context of decision-making during the battle, the soundness of a decision arises from the degree of situation awareness. Thus situation awareness is very important for commanders when leading troops in battle, because better situation awareness is a key to faster and better decisions (Endsley & Jones, 1997). That is because due to the amount of meaningful patterns and structures in the long-term memory of experts, they are able to achieve and sustain better situation awareness than novices (Sohn & Doane, 2004). This is the main reason why experts with less mental workload are able to comprehend and project the status of the situation better than novices (Endsley, 1995; Sohn & Doane, 2004). Thus only experts are able to decide quickly while using the advantages of naturalistic decision-making; novices on the other hand have to rely on analytical skills, and need more working memory and time than experts (Grier, 2012).

The structures or patterns that experts do have and novices do not are sometimes referred as tacit knowledge (Hedlund et al, 1998). Tacit knowledge is believed to be the type of knowledge that cannot be pulled out of the memory in words, such as for example knowing of what decision to make or how to do something that cannot be clearly voiced in a manner such that another person could extract and re-create that knowledge (Bennett & Bennett, 2014). In addition, tacit knowledge is also gained through implicit learning and once again is believed to be the basis for intuitive reasoning and decision-making (Patterson, Pierce, Bell & Klein, 2010). Implicit knowledge is knowledge stored in memory of which the individual is not immediately aware, but may be pulled up when triggered (associated) (Bennet, Bennet & Avedisian, 2015).

As the focus of the current doctoral thesis lies on small unit's commanders' decision-making in a battle context, there are other concepts that are believed to boost the performance. One such is the concept of unit cohesion, which is rooted in the battles of the 2nd World War and deals with the cohesion of German units (Shils & Janowitz, 1948). However, team related constructs such as different types of cohesion are not included in the concept of military cognitive readiness (Grier, 2012; Morrison & Fletcher, 2002). Grier (2012) still agrees with that team related constructs are essential part of successful military performance and recommends to examine those constructs in the future. A short overview of the construct called unit cohesion is given in the next section.

2.6. Unit cohesion and military decision-making

There are many concepts of cohesion, and in the military context, unit cohesion seems to be the one, which is used most commonly. It is the phenomenon that describes how and why the members of a certain military group develop and retain a pattern of behaviour that allows them to work together and achieve common goals, especially in stressful environments (Siebold, 2006). Unit cohesion is believed to be crucial in critical situations, where leaders must make quick decisions in a short period of time and relay them to the right people. However, there appears not to be strict consensus on what cohesion is – whether it is relationships among group members (including mutual trust and perceived military competence), a special form of motivation, or collective combat performance (Siebold, Crabb, Woodward & King, 2016). Different researchers have offered different models of cohesion in the military context that are not always mutually compatible (King, 2006; Kirke, 2009, 2010; Siebold, 2007) and recently Käihko's macro- and meso-level view on cohesion (Käihko, 2018; Käihko & Haldén, Käihko & Halden, 2020). The alternative approach to cohesion in the military context was recommended by Kirke (2009, 2010) and MacCoun, Kier and Belkin, 2006) stressing that cultural context of the organization should be taken into account and introducing the concepts of task cohesion and social cohesion.

Regardless of the ongoing debate, a standard model of military unit cohesion has been proposed and widely used (Salo & Siebold, 2007; Siebold, 2007; Salo and Siebold, 2008). Based on Siebold's (2007) standard model, unit cohesion consists of two broader categories: primary and secondary group cohesion. Both of these cohesion types further consist of two types of bonding. Primary group cohesion involves peer (or horizontal) and leader (or vertical) bonding, and secondary group cohesion involves organizational (immediate secondary group) and institutional bonding. Bonding refers to "the social relationship, both affective and instrumental, of changeable strength (weak to strong) between service members and their group, organization, and service institution" (Siebold, 2007, p. 288). Siebold (2007) claims that organizational bonding occurs on the lower e.g. company level, whereas institutional bonding relates to the members on the wider military level (the Army). Secondary group cohesion is more formal than primary group cohesion, interactions with secondary group members tend to be less intimate and mostly revolve around topics related to work.

Several studies have confirmed the association between group performance and unit cohesion (Beal, Cohen, Burke & MacLendon, 2003; Gully, Devine & Whitney, 1995; Mullen & Cooper, 1994; Oliver, Harman, Hoover, Hayes & Pandhi, 1999; Salo, 2006; Siebold, 2006). Accordingly, a widespread understanding exists that cohesion, especially primary group cohesion, is of critical importance in unit performance (Shils & Janowitz, 1948; Salo, 2006; Siebold 2006). According to some researchers, the most important component in a successful performance in critical battle situations is their superior's behaviour (Jacobs,1991; Mael & Alderks, 1993). Campbell (2006) has even stressed that the importance of other psychological factors is decreased if a leader can uphold group loyalty. Bartone and Kirkland (1991) have emphasised that the decisive factor differentiating cohesive, high-performance groups from mediocre groups is the behaviour of the group leader. However, unit cohesion could also have negative consequences such as demotivated and dehumanized behaviour of soldiers, depending on how unit cohesion was achieved (Pawiński, 2018).

It follows that unit cohesion and commanders' role in it should be good predictors of commanders' decision-making skills and performance during battle. As mentioned earlier, unit cohesion can be seen as a form of motivation, or at least unit cohesion seems to be the motivating factor behind the soldiers' will to fight in the battle. Grier (2012) has also considered motivation as an important component of military cognitive readiness. In the next section, a short overview of the concept of motivation is given.

2.7. Motivation and military decision-making

As mentioned earlier, motivation is considered an important category of military cognitive readiness (Grier, 2012). For the purposes of this doctoral thesis motivation is important, because it is considered to be a key component in differentiating between novices and experts and it leads to higher metacognition, which leads to sense making. Sense making in turn leads to expertise, which enables one to be successful in any field (Grier, 2012). Motivation consists of two important elements: locus of control and self-efficacy (Bandura, 1977). Locus of control can be divided into external and internal: an individual with external locus of control is more likely to hesitate in stressful conditions and most likely will not work hard to overcome obstacles, while an individual with internal locus of control is more likely to overcome obstacles, focus and accomplish the goals

(Thompson & McCreary, 2006). It is argued that good leadership can change locus of control from external to internal and vice versa (ibid).

Self efficacy can be understood as an individual's judgement of how well he or she can execute some course of action necessary for solving prospective situations (Bandura, 1982). Self-perception of efficacy, whether accurate or faulty, is also believed to influence thought patterns, choice of actions, and emotional arousal (Bandura, 1982). The higher the level of induced self-efficacy, the higher the performance accomplishments and the lower the emotional arousal (Bandura, 1982). Further, people with high self-efficacy tend to use meta-cognition while performing tasks (Coutinho, 2008). Locus of control and self-efficacy have an interactive effect on each other – individuals with internal locus of control tend to have higher self-efficacy and vice versa (Bandura, 1977). Self-efficacy thus seems to be based on metacognition and motivation (Grier, 2012).

One well known instrument used for measuring motivation is the Motivation at Work Scale (MAWS), which is based on a theoretical construct of self-determination theory (Gagné & Deci, 2005). The fundamental idea behind self-determination theory is that motivation is divided into two types: *intrinsic motivation* (derived from one's personal values and inner goals) and *extrinsic motivation* (instrumental motivation that arises from the tasks' perceived utility to some other task (Gagné et al, 2010; Ryan, Deci & Edward, 2000). Depending on the degree of internalization, extrinsic motivation spans from low self-determination to high self-determination and is divided into four subtypes (Ryan & Deci, 2000). Internalisation means the degree to which extent the goals and values of the task have been internalised, i.e. the degree to which one's goals, values and attitudes towards the task are regulated in a way that they match the goals and values of the task at hand (Gagné & Deci, 2005). The four subtypes are: (i) *external regulation* – doing an activity for the sake of obtaining awards of avoiding punishments; no goal internalisation is involved in this stage; (ii) *introjected internalisation* – regulation of behaviour through mechanisms of internal pressures to one's self-worth related to finishing the task, which is a low form of internalisation, when only the normative, inherently controlling aspects are internalised; (iii) *identified regulation* – action is motivated by one's identification with the general value or meaning of the task, and one has accepted the goal and value as one's own; and (iv) *integrated regulation* – one relates to the value of an activity to the point at which it becomes a part of one's subjective value system and habitual functioning.

It is important to note, though, that self-determination theory is not a stage theory and does not suppose that one must necessarily move through the four "stages" (Gagné & Deci, 2005). Instead, Deci and Ryan (2000) highlight the importance of autonomy, namely the autonomous regulation of extrinsic motivation could also lead to positive psychological as well as performance outcomes. The role of autonomy is emphasized even more by Kusurkar, Ten Cate, Vos, Westers and Croiset (2012) who proposed that identified regulation and intrinsic motivation can both be seen as an autonomous type of motivation and showed that autonomous motivation positively affected academic performance, but through a deep strategy towards study and higher study effort. Moreover, Buch, Säfvenbom and Boe (2016) have integrated self-determination theory and self-efficacy approach and showed that the effect of self-efficacy on perceived military performance depends on the level of intrinsic motivation. Karton, Männiste, Tepp and Kornilov (2018) found that intrinsic motivation significantly predicted conscripts' desire to continue their career in active service.

However, motivation and perceived self-efficacy are not always sufficient predictors of better performance. A meta-analysis by Stajkovic and Luthans (1998) found that the effect of self-efficacy on task performance is mediated by the complexity of the task: the higher the task complexity, the weaker the relationship between self-efficacy and performance. Therefore, it is reasonable that in highly complex tasks, such as decision-making in battle like situations, the influence of self-efficacy on performance depends on previous knowledge and experience, e.g. experience in the field.

2.8. Measuring military decision-making in the battle context

In order to construct an optimal instrument for measuring decision-making skills in critical, high-stakes battle-like situations, it is important to have a knowledge of previous work in the field. For this purpose, as a part of this study, a systematic literature review was conducted. In the literature review, a thorough analysis of current instruments for measuring decision-making skills in the military context was conducted. For the purposes of the current research, the interest lay in a test that is cost-efficient, can be conducted in a classroom setting within 1–2 hours ideally, and is easily adaptable to different kinds of scenarios. For more considerations, see Article **II**. In the course of the literature review, it was discovered that measuring low-level military commander's decision-making skills in a battle leading context can be broadly divided into three categories (see Article **II**):

(i) Live performance;
(ii) Simulated performance;
(iii) Tested performance.

According to the literature review (Article **II**), live performance type of measures are intended to mimic a real life situation as closely as possible and they require setting the situation up from the ground, as well as participants actually performing in this scenario. Examples of these kinds of tests are staging a live burn, which means setting fire to a house replica and asking volunteers to fulfil the task in this environment (Cohen-Hatton & Honey, 2015), having the participants trace a quarry in wilderness (Spiker & Johnston, 2013), or observing participants in training courses set by their employing institutions (Thomas & Hirschfeld, 2015). As our literature review (Article **II**) showed, live performance type of measuring always requires constructing a unique setting for any specific test, which is costly and time-consuming. Live performance tests are usually halted or paused at either fixed or random intervals to administer different type of questionnaires to the participants. They are typically asked to answer questions such as why are they doing what they do, what was generally going on in the situation just before pausing, what were the events of decisions leading to the current outcome and what they think of the current situation. In addition to self-report questionnaires, objective data is also typically collected in the case of live performance tests. For example, psychometric instruments can be used for data gathering, such as nerve sensors and heart rate monitors (Article **II**).

Simulated performance tests are intrinsically very similar to the live performance tests. The major difference between the two is the construction of scenarios used for testing. Whilst live performance takes place in actual settings, simulated performance tests require an artificial or virtual simulation of the situation. For example, participants may be tasked with clearing a corridor of enemy units (Hale, Stanney & Malone, 2009), but instead of an actual physical corridor, a virtual reality headset is given to participants. As in the case of live performance, it is also common to share questionnaires during, before or after testing. In the case of simulated performance, additional psychometric measuring (nerve sensors, heart rate monitors etc.) is also widely used (Article **II**).

The tested performance type of tests differ from the previous two types in the way that no actual performance is carried out during the testing. Instead, hypothetical *what-if* scenarios are administered to participants and they are then asked to either solve or identify the problem-situation in the scenario. This can be done either by participants listing all the relevant incidents they can see in the battle situation (Lussier & Shadrick, 2004), and/or by ranking presented solutions in the order of their suitability. Compared to other types, tested performance type of tests are more cost-efficient, as they do not require constructing an actual physical environment or purchasing an expensive simulator (Article **II**). Once compiled, tested performance tests are reusable and easily adaptable to different scenarios.

One particular type of tested performance test appears to be the Situational Judgment Test (SJT). In SJT, test-takers are presented with a variety of work-related situations and then asked what they should (or would) do in each situation (Champion, Ployhart & MacKenzie, 2014). SJTs have been a popular tool for personnel selection since 1873 (Weekley, Ployhart & Holtz, 2006), but it has been argued that research on SJT began in 1990 with Motowildo and his colleagues (Motowildo, Dunnette & Carter, 1990) who asserted that SJTs emanate from the tenet of behavioural consistency (i.e. past behaviour is the best predictor of future behaviour).

Since then the research into SJTs has been quite extensive and it has been found that content of SJTs can be developed in a way that test taking performance (i.e., test scores) reflect not only past behaviour, but a combination of many different constructs or composite competencies (Chan & Schmitt, 2005). For example, Christian, Edwards and Bradley (2010) has proposed a comprehensive typology of constructs that were measured using their procedure: applied social skills, heterogeneous composites, leadership, basic personality tendencies, job knowledge and skills, and teamwork skills.

However, there are other findings, based on which it is reasonable to conclude that SJT responses can also be expected to be only a function of generic and domain-specific job knowledge gained through experience or formal education (Ployhart & Weekley, 2006). This argument seems to be supported by Johnson and Oswald (2010), who state that SJT content can reflect a wide variety of constructs, but empirical analyses usually result in a single general situational judgment construct.

SJTs can be developed in a way that scenario and situations can be presented in paper format or in audiovisual format. It has been found, for example, that test takers find it more interesting and motivating if the situations are presented in audiovisual format (Richman-Hirsch, Olson-Buchanan & Drasgow, 2000). In such a format, they find themselves more involved and thereby acquire deeper learning and achieve learning objectives better (Tan, Tse & Chung, 2010). Many previous studies also have proved that SJTs reliably predict work performance (Hauenstein, Findlay & McDonald, 2010; McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001; Saldago, Viesewaran & Ones, 2001; McDaniel, Hartman, Whetzel & Grubb, 2007). In addition, in reviewing validity evidence related to SJTs, Whetzel and McDaniel (2009) reported, based on individual and meta-analytic studies, that SJTs exhibited good criterion and construct validity. Thus the most suitable type of test for further consideration and development was considered to be SJT.

3. METHODOLOGY

The current doctoral thesis comprises four distinctive phases, which have been carried out for achieving the set aims (see also Figure 1):

(i) Phase 1. Preparation for instrument development (covered in Article **I** and Article **II**)

(ii) Phase 2. Instrument development, data collection, instrument quality check and improvement based on empirical data (covered in Article **III** and Article **IV**).

(iii) Phase 3. Describing the results of test based on different groups (covered in Article **IV**).

(iv) Phase 4. Finding out factors which predict better results in SJT (covered in Article **V**).

3.1. Phase 1: Preparation for instrument development

This phase consists of two clearly distinguishable steps and was meant to answer the research questions 1 and 2. Firstly, theoretical overview about the meaning of the problem and special attributes of problem solving in the military sphere was given (in Article **I**). This step gave insight into what type of problem situations to include for instrument development phase – e.g. it served the purpose of identifying the basis for scenario development and test structure. It also gave some theoretical considerations, which helped to clarify the role of problem solver in a very specific military sphere. Secondly, a systematic literature review was undertaken in order to find out and give an overview of the instruments that have been used for measuring or/and predicting decision-making skills in low level (tactical) battle leading context (Article **II**). Step two helped to find out the existing instruments, in order to adapt or develop an instrument suitable for the purposes of this study.

In step one of phase 1, semi-structured interviews with long-term servicemen were selected as method for gathering data. It has been found to be suitable method, when there is a need for new information, which researcher does not know yet (Lincoln & Guba, 1985). Considering the fact that the interviewer and all interviewees were experienced military officers, the interview was conducted in the form of dialogue between equal partners (Kvale, 2006). Before the interview, the procedure and ethical points were introduced to the interviewees (Cohen, 2007). During the interview, interviewees were asked questions to which they were able to answer freely. They were also encouraged to clarify questions, illustrate and visualize their answers if needed and recommended to think aloud.

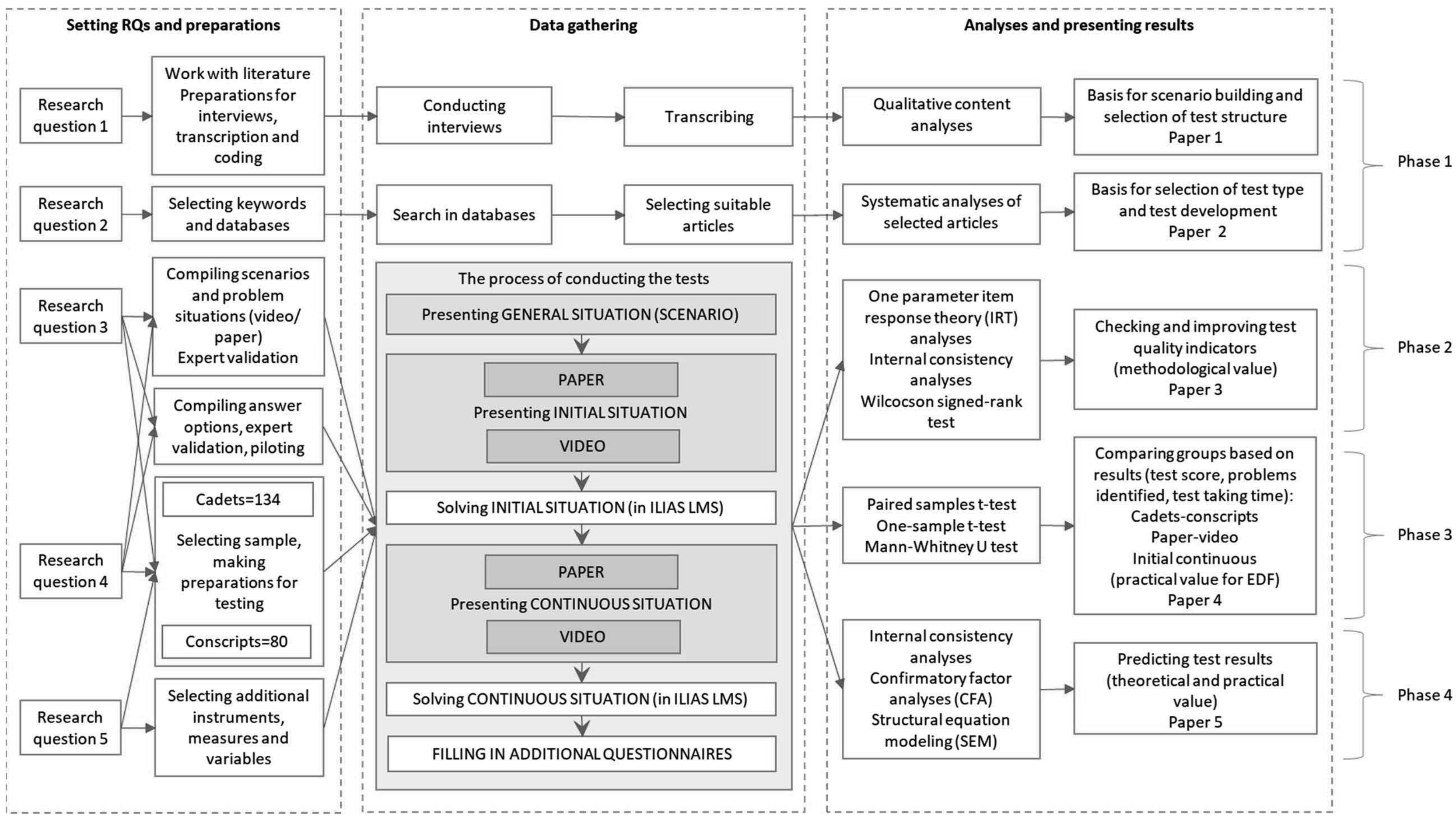

Figure 1. Research design.

Interviews were conducted with six experienced military officers (4 senior officers, one junior officer and one non-commissioned officer). The criteria for selecting sample were a long service period in EDF or Estonian Defence League. A convenient sampling method was used, meaning that the service positions made it easier to organise interviews. Two interviewees had 23 years, one 21 years, two 19 years and one seven years of previous service experience in different positions in EDF including previous experience in the field of military training. All but one (31) were older than 37 years.

The data gathered with the interviews were transcribed and analysed using the program NVIVO 10. A targeted deductive approach was preferred, which means that only information of interest to the researcher was looked for (see more in Article **I**).

In step two of phase 1a, a systematic review of literature (see Figure 2) was undertaken following PRISMA guidelines (Moher, Liberati, Tetzlaff, Altman & PRISMA, 2009). It allowed to determine whether an instrument of interest to the current study already exists, or whether there is a need to develop one (Article **II**).

The literature search was broken down into four phases: firstly, identifying potentially relevant papers in the EBSCOhost Web database; secondly, assessing the records identified via database search at title level; thirdly, assessing remaining papers at abstract level; and finally, assessing remaining papers at full text level. Papers were identified via EBSCOhost Web service. The following databases were selected for identifying relevant records for the literature review: Academic Search Complete, American Doctoral Dissertations, Central and Eastern European Academic Source, CINAHL Complete, eBook Collection, E-journals, ERIC, Health Source: Nursing/Academic Edition, Master FILE Premier, Match SciNet via EBSCOhost, MEDLINE, MLE Directory of Periodicals, MLE International Bibliography, PsycARTICLES, PsychINFO, and Teacher Reference Center. Only papers written in English were considered. The search terms are listed in Article **II**. Initially, 986 records were detected.

The criteria for including studies in this review consisted of five questions: (i) Is the topic of the study decision-making in critical situations? (ii) Does the study focus on measuring instruments? (iii) Does the study focus on training decision-making skills? (iv) Is the study related to the military sphere? (v) Are the methods used in the study quantitative? The studies were first screened at title level for suitability, the scale applied was: yes-maybe-no. The process of screening was as follows: first, two researchers independently read the titles of the articles and categorized them as "suitable for abstract level screening" or "unsuitable for abstract level screening". The minimum level of agreement between researchers' opinions was placed at 80%. If the level of agreement was less than 80%, the results were discussed and the process was repeated; otherwise, the researchers proceeded to the next phase. This process was repeated at each phase (see Figure 2).

Eight criteria were used to describe how appropriate an instrument was for measuring decision-making skills in the military context: (i) Related to the military sphere; (II) Sample: military personnel; (III) Type of performance: sub-indicators: (iii.a) live performance, (iii.b) simulated performance, and (iii.c) tested performance; (iv) Applicable to military sphere; (v) Used in training; (vi) decision-making in critical situations; (vii) Level of war; (viii) Reported quality indicators. For clarity, the scoring was kept as simple as possible: 0 – no, 1 – so-so, 2 – yes for indicators other than (vii) and (viii), and 0 – level of war not presented, 1 – operational or strategic level, 2 – tactical level for (vii). In the case of (viii), the sum of quality indicators gave the final score: 0 –quality not reported, 1 – quality reported. Quality indicators were categorized: MI: main instrument's quality, AI: additional instrument's quality, R: result's quality. For (viii) it was possible to get 3 points altogether. Finally, all the studies were ranked on the basis of the indicators introduced in this chapter.

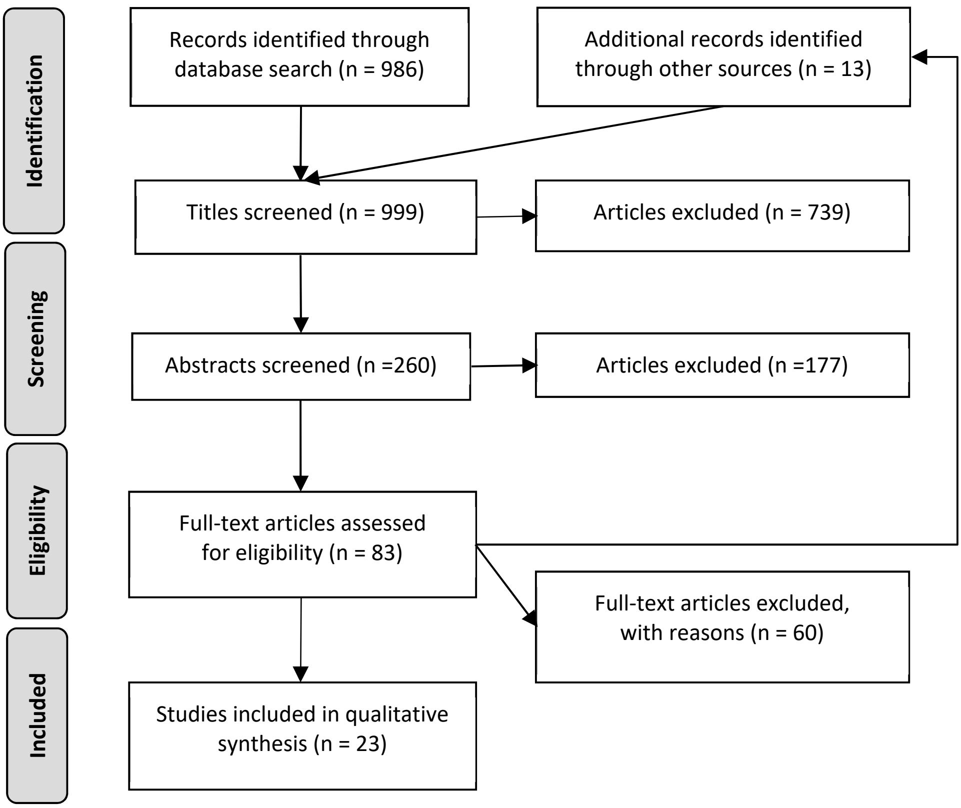

Figure 2. Flowchart of the literature search procedure (Article **II**).

3.2. Phase 2: Instrument development, data collection, instrument quality check and improvement based on empirical data

This phase consists of three clearly distinguishable steps and was meant to answer research question 3. First, based on the results of previous phase (Article **I** and **II**), an instrument for measuring low level military commanders' decision-making skills in battle-like situations was developed. The process of developing that instrument is briefly described in Article **III** and more in detail in Article **IV**. Secondly, the sample was selected and data collection was carried out, which was again briefly described in Article **III** and more in detail in Article **IV**. Finally, based on the empirical (collected) data, the quality of developed instrument was assessed and improved. This process was described in Article **III**.

The situational judgement tests measuring decision-making skills (SJTDM) developed in the current study were based on the structure of strategic performance problems described by Jonassen (2000) because this type of problem appears to be the closest to problems in a battle situation (see more in Article **I**). This type of problem involves real-time, complex and integrated activity structures where the performers use a number of tactics to meet a more complex and ill-structured strategy while maintaining situational awareness (ibid). Thus, based on Article **I**, it was decided that the SJT was developed as a two-part problem situation where the first part (initial situation) started the situation and involved simpler incidents, and in the second part the situation continued and escalated into more complex incidents.

Developing general scenario and situations

Initial scenarios for further development were selected amongst Tactical Decision Games, which were created by advanced officer course participants during a pedagogy course, which at the time was run by the author of the current doctoral thesis. The further compilation of the scenarios and answer options were also conducted by the author. All scenarios and answer options were then reviewed by a junior officer (rank captain), who had previous experience as platoon/ company commander for four years, one year of experience in teaching platoon tactics and one year of experience as a cadet course commander.

When developing a scenario, the principles proposed by Salas, Priest, Wilson and Burke (2006) were followed: (1) determine the objective and the level of difficulty, (2) the scenario should be based on realistic practices, representing what the soldiers may actually experience on the battlefield (ibid). Scenarios were drawn up on defensive activities (hasty defence and delay) at the infantry platoon command level. The build-up of the scenarios included the need to understand the given information (incidents etc.) and decide in order to solve tactical problem situations. The scenario was divided into two parts: 1) the general situation; 2) the initial situation and the continuous situation (in this doctoral thesis as well as in all articles, the term "continuous situation" is used, which can be taken to mean a follow up situation).

The general situation was set up in a way that it required decision-making to solve the situations presented. Consequently, the structure of general scenario included listed attributes: (i) role of the solver; (ii) list of equipment, armament and manpower available; (iii) call signs of superior commanders, subordinates and other needed units; (iv) description of the general situation including time, weather and terrain; (v) higher commander's intent and orders, tasks of own and neighbouring units; and (vi) detailed description of recent incidents, separating out all the details that the platoon commander sees (enemy activities related to the landscape and time etc.) and hears (battle voices, radio sessions etc.) in a situation.

Two scenarios (hasty defence and delay) that were selected were further developed and prepared so that each of them consisted of two ongoing continuous situations: initial (easier) situation, which escalated into continuous (more complex) situation. While developing the situations, the principles proposed by Weekley, Ployhart and Holtz (2006) and Legree and Psotka (2006) were followed: presented situations should include critical events, which emphasize specific competences or general knowledge in the particular field. Consequently, both situations presented the most recent incidents that a platoon leader typically faces in the battlefield and that led to the need for decision-making. After development, both scenarios (hasty defence and delay) consisted of the following elements: (i) general scenario description (PowerPoint presentation and written description); (ii) initial situation (video and paper format); (iii) test questions (open question and response options) of the initial situation; (iv) continuous situation (video and paper format); and (v) test questions (open question and response options) of the continuous situation (see https://sisu.ut.ee/sjt).

Two different ways of presenting the situations were selected and accordingly two types of presentation means were prepared: 1) video version and 2) paper version. The reasons are discussed more detail in Articles **III** and **IV**.

Video versions of the situations were developed first. The paper versions of the situations were made after the completion of the video editions, since it was possible to see the already finished video and use it to write a paper version. In doing so, it was possible to accurately describe the video in order to achieve comparability of two versions. As a result, both versions were identical in content, meaning that everything that was visible and audible in video format was written down in detail in paper format.

Preparing video based presentations of the situations

The videos were made using the EDF licensed software Virtual Battlespace 2 (VBS 2) with the help of the specialists of ENDC Simulation Centre. First, a suitable map of the terrain (including existing roads, settlements, rivers, elevations, etc.) for each scenario was selected, which was subsequently designed to match the scenario (for example, trees, shrubs, etc. were added to the existing map). Then, all units involved in a particular situation that needed to be visible to the test takers (e.g., the enemy's as well as own troops) were placed on the map. After that, a plan for troop movement and fires was created (for example, enemy moves on a specific route for some time, looks at us, discovers our units, hides and opens fire), and the animation script was made on the basis of the constructed plan. Once the unit movement plan was in place, a video recording plan was made. When everything was ready, the simulation script was run and a pre-planned 3–5-minute-long video clip of the simulation was recorded. After that, radio sessions were recorded as separate audio files and added to video clips using the Movie Maker programme. The process of making videos is described in more details in Articles **III** and **IV**.

Making paper-based presentations of the situations

The paper versions of the situations were made after the completion of video versions, since it was possible to see the ready-made video and use it for writing a paper version. In doing so, the actions in video were described as accurately as possible.

Developing answer options and selecting the SJT response method

All the answer options were developed in Estonian language, however, the English translation is available online (see https://sisu.ut.ee/sjt). There were two types of answer options: open answer options and multiple choice answer options. Open answer options asked test takers to list all of the problems they discovered in the given situation. This was similar in all four situations.

The multiple choice answer options were compiled based on behaviour based response instruction because such items tend to evoke a more behavioural response and have shown more favourable characteristics than knowledge based response instruction (Ployhart & Ehrhart, 2003). While compiling multiple choice answer options, the aim was to achieve credibility and comply with real-world rules. During the process, a description of the best answer option for the situation was formulated first, followed by the description of the worst solution. Then, other answer options were developed by adding or omitting details so that finally all the answer options to a particular situation would constitute a continuum of better and worse answer options to accompany the problem situation at hand, as proposed by Bergman, Drasgow, Donovan, Henning and Juraska (2006).

In compiling answer options, attempts were made to avoid descriptions that would make them appear obviously correct or false. This was to minimise the possibility of deception or just guessing by the test takers. Finally, to some extent, each answer option reflected nearly similar actions, but sometimes some actions were added that would turn the otherwise good answer option into a bad one. In terms of the amount of text, the answer options remained more or less equal to avoid a situation where better solutions are longer and thus pose a visually more obvious choice for respondents.

At the beginning, there were 12 answer options developed for every situation (two per each of six difficulty level), but the number was reduced to six during the keying process. In the final version, there were thus six multiple choice answer options created for every situation, altogether there were 24 answer options. A six-point scale was used to assess the quality of each of the answer option (1 – very poor; 2 – poor; 3 – rather poor; 4 – rather good; 5 – good; 6 – very good). The commonly used rank-all response method was applied (Campion, Ployhart & MacKenzie, 2014) in the case of SJTs. In other words, respondents were instructed to indicate their preference for supplied actions according to some characteristics (i.e. what they feel is the best course of action, second best and so on). The respondents were instructed to indicate their preference by ranking all answer options. Each number was allowed to be used once only. The rank-all method was chosen because it appears to be more distinctive, reliable and valid and the likelihood of getting a good result from a random selection is reduced (Weekley, Ployhart & Holtz, 2006). Rank-all method is also preferable because it makes it possible to avoid certain problems that arise with Likert-type scales, such as acquiescence responding (Cheung & Chan, 2002) choosing between similarly attractive options (Brown, 2012), and "halo" effects (Bartram, 2007). It also helps to avoid scale range errors and errors caused by a different response style, which some researchers have dealt with in their studies (Hedlund et al, 1998).

Developing a scoring key for SJT

An expert-based method was chosen to develop the scoring key, because it appears to be the most commonly used method for key development (Campion, Ployhart & MacKenzie, 2014). In the case of an expert-based method, it is possible to use either expert consensus or score averaging or both combined (McDaniel & Nguyen, 2001) as it was done in the case of current doctoral thesis. At first, the compliance of pre-compiled (by author) and expert opinions about the level of correctness of answer options was determined. For this, an expert group of 14 persons was formed and the keying was carried out from February to April 2017.

The expert group consisted of:

- Six Officers and Non-Commissioned Officers of the NDC Chair of Tactics, Infantry Platoon Tactics Group;
- Four Officers from the NDC Officers School;
- Four Officers (with infantry background) from the NDC Officers School.

All experts were contacted by e-mail or phone and the time and place were agreed on. The key development process took place in a computer lab where it was possible to introduce the scenarios using a multimedia projector and show video clips. Experts were provided with the written versions of the scenarios (descriptions of the situations on paper together with maps with added military graphics). In addition, computers equipped with headphones were available to the experts, which enabled them to review the videos as many times as they wanted.

The procedure of keying with selected experts was as follows:

- Firstly, the scenario was introduced (using Power Point presentation). Printed handouts were also provided to the experts (which consisted of the description of the situation together with maps and radio transmissions used).
- A video of each situation was then shown. It was possible for experts to watch the video independently, if they wished to do so.
- After viewing the video, the experts were provided with worksheets and asked to individually rank all answer options on a scale of 1 to 6, where 1 = very bad, 2 = rather bad, etc. They were also told that at this stage, each number should be used twice (at the beginning there were two options corresponding each difficulty level). Each worksheet consisted of possible answer options (marked with capital letters A-L). Experts did not see the tentative scores assigned to each solution option by the authors (solutions were mixed in advance and sequenced randomly).
- Experts were then divided into groups of three in a way that if possible, both the officers and non-commissioned officers would remain in the group. The same assignment was then completed in the group. In doing so, experts were encouraged to discuss within the groups and reach the consensus about the sequence of answer options. In doing so, experts had the opportunity to use individual worksheets they had filled in earlier.
- When that was done, both the individual and group worksheets were collected from the experts. Then, recommendations to improve the scenarios, situations or solutions were asked from the expert group (for example better formulation of sentences).
- Finally, experts were also asked if they found it necessary to make recommendations for reviewing the answer options in order to improve their reality, variability and compliance with the overall scenario and situation as recommended by Weekley, Ployhart & Holtz (2006).

After experts filled out their documents, the following procedure was undertaken by researchers in order to choose the optimal six answer options out of 12 and to determine the level of correctness of those six answer options for each situation:

- The mean and median of all experts' estimates were calculated.
- The best suited set of six answer options was then selected and their level of correctness calculated by comparing the results of the experts' estimates.
- In the expert estimates about the level of correctness, similarities between expert estimates were looked at, with the criterion of one point up or down compared to the average estimate. Attention was also paid to how well expert ratings for the answer options matched the initial ratings by the researchers.
- Experts whose correctness estimate given to a particular answer option was bigger than one point up or down compared to the summarised average (individual and group) were later questioned to determine the reasons for the differences. If, in the experts' view, there was a misunderstanding, the actual meaning of the content of the answer options was explained to the experts in each particular case and they were given the opportunity to change their estimates. If necessary, the answer options were corrected.

Principles for evaluating test results

The results of multiple choice answer options and open answer options were evaluated. In addition, data about the test completion time were automatically saved in the e-learning environment.

The scoring schema used in assessing the performance of matching tests is shown in Table 1. The results of the test takers were compared to the expert group, which means that when the test taker gave the same number to the same answer as the expert group, he got the maximum possible amount of points. Every answer that differed from that of the expert group consensus gave the test takers fewer points, so the more the participant was wrong, the fewer points he received (Table 2). Thus the higher the final score was, the better the result.

Table 1. The schema by which points were given in the case of matching tests (Article **IV**).

		Key based on expert group					
		1	**2**	**3**	**4**	**5**	**6**
Test takers' choice	**1**	6	5	4	3	2	1
	2	5	6	5	4	3	2
	3	4	5	6	5	4	3
	4	3	4	5	6	5	4
	5	2	3	4	5	6	5
	6	1	2	3	4	5	6

Open responses meant that test takers had to formulate answers in their own words and write down all the problems they were able to discover in the situation presented to them. In order to assess the open answers, numerical values were assigned to each of the answers. It was done by two officers related to the study on the basis of previously agreed criteria and procedures. The agreed procedures were as follows (Article **IV**):

1) Both evaluators at first read and rated the participants' responses independently.
2) Then, the evaluators compared their results, and if more than a two-point mismatch appeared, the possible reasons were discussed and consensus was found.
3) Finally, the arithmetic mean of the points given by both evaluators was calculated.

The agreed criteria of coding open responses were as follows:

1) No point was given when test taker found units, equipment, armament and landscape inadequate or non-compliant for completing the task (unless the impact on its own unit was mentioned or analysed in the statement).

2) No point was given, if the test taker described persons' (who have some role in the situation) behaviour in the scenario as inappropriate or inadequate.
3) No point was given for pointing out false or completely non-essential aspects of the situation.
4) No point was given when a test taker listed the opponent's equipment (given in the scenario), without identifying their potential impact on its own unit.
5) No point was given in tha case of continuous situation when a test taker referred to something that had already been pointed out while filling out an open response in an initial situation (this was due to the fact that both scenarios had initial situation and continuous situation, which meant that there were similarities).
6) A point was given for mentioning each item considered important by the test developers and the expert group.
7) A point was given for mentioning any item that the test developers had not considered to be of direct relevance to the scenario, but which, however, was correct in the context of the situation, and indicated how deeply the respondent was able to understand the situation.
8) The spelling errors, the writing style and the correctness of the terms were not considered important if the meaning was understandable.
9) The composition and length of the sentences was not considered important. When two essential items were clearly indicated in one sentence, then a point was given for both of them.

In step two of phase 2, the detailed procedure of selecting the sample and conducting the tests (gathering the data) was undertaken. The sample of the current doctoral thesis consisted of cadets of the Estonian National Defence College (ENDCOL) and pre-call conscripts of Reserve Platoon Leader Course. The reason for this was to have more experienced/advanced (cadets) and less experienced/advanced (conscripts) participants in the sample. The sample consisted of all of the land forces basic course cadets who were studying at ENDCOL in 2017. This means that all of the different possible experience levels were presented in the case of cadets – the oldest cadet course had three years of military studies and training behind them and the youngest course had just entered studies without having much previous military training except for conscription in the EDF. The overall number of cadets who studied in the ENDCOL during this period (population) was 146 and all of them were asked to join. The final number of cadets who completed the tests was 134 (91.6%). Twelve cadets could not participate; of those two were ill and 10 had other reasons for not participating. The overall number of conscripts was 81 and again, all of them were asked to participate. Only one conscript could not participate, thus the final number was 80 conscripts (see Table 2).

Prior to testing, the sample was divided into two groups as equally as possible. One group conducted the test in video version, and the second group in paper version. The process of dividing the sample into two groups (Articles **III** and **IV**) was carried out as described in the next section.

Table 2. Descriptive data of the sample (Article **III**).

Group	Population	Sample	Partici-pants (%)	Age	Education %		Male %
					Secondary	Higher	
Cadets	146	134	91.6	23.7	94	6	96%
Conscripts	81	80	98.7	20.9	86	14	100%
All	227	214	94.3	22.3	91	9	98%

The process of dividing participants into two groups (applicable for the rest of the phases of this doctoral thesis)

Prior to testing, the sample was randomly divided into two equally sized groups. One group conducted the test in video version, the second group in paper version. The basis for dividing the cadets into two groups was stratified sampling (based on ranks) combined with random sampling. In addition to the ranks, previous military service before the commencement of studies was also taken into account. Firstly, every cadet course was divided into two groups. The first group consisted of those cadets who had after mandatory conscription been involved with active military service before the commencement of studies. The second group consisted of cadets who came to study only with mandatory conscription experience, thus without previous active military service experience. Then, the two groups were further divided randomly into two equal sized groups so that both groups had the same amount of more experienced and less experienced cadets. Subsequently, both groups were divided into three subgroups based on military ranks: (i) cadets without rank; (ii) junior non-commissioned officers; (iii) senior non-commissioned officers and officers. The Microsoft Excel random selection function was used to divide each of these sub-groups randomly into test groups. In order to form equal groups, both groups were then merged so that both groups had an equal amount of experienced and ranked members. In both groups, the first half of the group formed the paper group (67); the other half, the video group (67).

In the case of the conscripts, dividing into groups was done by the managing body of the reserve officers' course, who organised and handed out testing schedules to participants. Thus conscripts were able to participate according to their pre-prepared weekly and daily schedule. However, since the standard basis of group formation in the reserve officers' course is based on the principle of equality, it is reasonable to assume that the sampling was basically random. Regardless of the previous distribution, many conscripts could not participate in the groups assigned to them for different reasons, so they were able to participate in the different day and different type of test. Because of this, conscripts were divided into groups unequally: the paper group was a bit larger (45) than video group (35). One conscript out of 81 did not participate in the tests for unknown reasons, but as participation was voluntary, being absent probably meant unwillingness to participate.

The process of gathering the data (conducting the tests) (applicable for the rest of the phases of this doctoral thesis)

Tests were carried out in computer classrooms in groups (video and paper) of varying size (typical size of a group was approximately 20 persons). Two researchers (testers) were present at any time in both groups. One of them was involved in introducing scenarios as well as having control over the activities of participants all the time. The other one was involved in making notes about students' activities in the classroom by fixing exact times of specific events (for example exact times, when scenario presentation started and finished etc.), as well as helping participants in resolving issues that arose in the course of the testing. The procedure for conducting tests (gathering data) was as follows (Articles **III–V**):

Introduction of the test to the participants

First, there was an introductory presentation which revealed the necessity, reasons and procedures for conducting the tests (including the use of the ILIAS e-learning environment). In addition, the ethical aspects of the study were introduced to the participants, including how the anonymity of the participants was ensured. Participants were also informed about the sampling principles and the possibility of receiving personalized feedback on test results. Finally, participants were asked to take both the tests and questionnaires seriously.

The process of conducting the tests

1. PowerPoint presentation for guiding work through the scenario. There were two presentations: one for providing instructions on the video-based test and another one on the paper-based test. In both presentations, the first five slides were for introducing the general scenario and the next slides were different, depending on the version of the test. In the notes of the presentation, it was specified which slides to show along with the following phases (2–5) in conducting the Situational Judgement Test. There were also some guidelines in the notes to explain what the people organizing the test should do. The general scenario provided participants with an overview of their role, the available resources, the structure of the unit, necessary call signs, the environment, the task of the higher command level, and the task of his own unit. After presenting the situation, participants were given the opportunity to ask questions. This stage was conducted twice, first for hasty defence tests and then for delay tests.
2. Presenting the initial situation. The initial (first) situation revealed the latest events that had happened and thus led to the first situation in which respondents had to decide (the first problem that had to be solved by respondents). This was done by showing an approximately four-minute long videos (the group members were able to see and hear everything) or handing out paper-based illustrated descriptions and graphics (the group members were able to read the

information from documents). The members of video group were told that they would be able to watch the video only once and in addition to just remembering, they were advised to take notes (which they could use later while solving the situation) while watching the video. The paper group was told to take the time they needed to get familiar with the situation, and they were also told that that they could use all the documents later while solving the situation.

3. Solving the initial (first) situation. This stage was conducted in the e-learning environment ILIAS, where both groups solved the situation exactly the same way. Solving the situation consisted of two parts: open answer options and matching answer options. The open answer options were to be entered first; participants had to write down the problems they were able to identify in the situation. Then they had to solve the matching answer options by matching six response options to six digits. The answer options were presented in random order. During the test, the members of the video group had to rely upon their memory or the notes they had made earlier. The members of the paper group were allowed to browse the documents they were given earlier. There was no time limit set for resolving the situation.
4. Presenting the continuous situation. This stage was conducted like the 2nd stage. Respondents received an update of the situation, which was the continuation of the previous situation, but more difficult this time.
5. Solving the continuous (second) situation. This stage was conducted like the 3rd stage.

After that, stages 1 to 5 were repeated with the second scenario. See more about conducting tests in Articles **III** and **IV**.

The process of assessing and improving the quality of developed SJTDMs (applicable for the rest of the phases of this doctoral thesis)

The purpose of step three of phase 2 was to assess the quality of developed instrument and improve it if needed. Firstly, based on gathered data (test scores) and using IBM SPSS Statistics version 25, internal consistency (alpha) of the SJTDM was assessed. It was done by evaluating all of the 24 multiple choice answer options together. The acceptable reliability score should be between .70 and .90 in accordance with Nunnally (1978). Then, for improving the quality indicators of the SJTDMs, one-parameter Item Response Theory (IRT) analyses were conducted using the WinSteps 4.0.1 program. One-parameter IRT analysis was chosen, because it is recommended in a case where items have different levels of difficulty (typical in the case of competence tests). The IRT analysis allows the quality of each test item to be evaluated, which makes it especially suitable for the process of instrument development. In short, IRT helps to identify test items that are illogical and adjust them by changing the scoring schema (Article **III**). During IRT analysis, four quality indicators of SJTDMs were monitored:

1. The goodness of the scoring key: it was checked on the test level whether the sequence of answer options estimated by experts during the process of keying was in accordance with the empirical data. It was also checked whether higher scores were given to options that were also more difficult to answer and whether six levels of correctness for each test is an optimal scoring key (if so many levels could be empirically differentiated). Ability mean provided by IRT analyses was used as a basis for evaluating the goodness of scoring key. When the ability mean of the close answer options was similar and the number of respondents of each answer option was small (5% or less), these answer options were merged. In other words the same amount of points was awarded to each of these options while re-scoring the answers. In addition, this change in scoring was also justified through analysing the content of the answer options: when the change was not reasonable based on the logic of the content on answer options, then the change recommended by empirical number was omitted.
2. The goodness of test items in discriminating the respondents: it was assessed how well different test items distinguished between the respondents. An estimate of discrimination measure (because the discrimination measure is set to 1.0 in case of 1PL IRT model and only calculated in case of 2PL IRT model) provided by IRT analyses was used for this, with the suggested values set between 0.5 and 2.0 (1.0 is considered ideal discrimination) as suggested in the WinSteps manual
(see http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm).
3. The goodness of item fit of the test items: it was assessed how well the scores of each item correspond to the expected difficulty level of these items in the test. This was measured by a correlation coefficient provided by the IRT analyses; the threshold of .20 was considered as acceptable as suggested in the WinSteps manual
(see http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm).
4. The variation of the difficulty measure of the test items. This measure shows how difficult the test items are on a scale where 0 is set in the middle. For a good measurement, it would be important to have some test items placed around the middle of the scale, but also to have items that are simpler or more difficult (away from the middle of the scale).

Finally, internal consistency of the SJTDMs was assessed, with the acceptable reliability score expected between .70 and .90 (Nunnally, 1978).

3.3. Phase 3: Describing the results of the SJTDM based on different groups

Phase 3 was meant to answer research question 4: how accurately does the instrument distinguish novices from experts/professionals in a simulated platoon leader battle scenario in the example of EDF cadets (advanced) and conscripts (novices)?

This was the only phase of the study where all the collected test data (including test scores, the number of problems identified and test taking time) was used for group differences analyses (see more details in Article **IV**).

The objectives of phase 3 were:

1. To identify whether the results to initial situation and continuous situation differ across the whole sample and different groups (cadets/conscripts and paper/video).
2. To identify whether the number of problems identified in situations differ across the whole sample and different groups (cadets/conscripts and paper/video).
3. To identify whether the test taking time differ across the whole sample and different groups (cadets/conscripts and paper/video).

The sample used in this phase was the same as in phase 2 (see Table 2). The SJTDM used for data gathering and the procedures used were also the same as developed in phase 2 (see more about sample, instrument development and quality improvement also in Article **III** and Article **IV**). The data used in analyses had previously passed the process of quality improvement (see step 3 of phase 2) based on IRT analyses.

Both dependent samples and independent samples analyses were undertaken to identify group differences. All the analyses in phase 3 were conducted by using IBM SPSS Statistics version 25. First, the Kolmogorov-Smirnov normality distribution analyses were conducted. Thereafter, dependent samples and independent samples analyses were conducted. In the case of dependent samples analyses, the Wilcoxon signed-rank test was used. In the case of independent samples analyses, t-tests were preferred if the prerequisites were met (checked with the Levene test). If the t-test prerequisites were not met, Mann-Whitney U-test was used. The statistical significance level of the analyses was chosen to be $\alpha = .05$.

3.4. Phase 4: Finding factors predicting results of SJTDM

Phase 4 was meant to answer research question 5: what are the attributes that predict a better result in decision-making test based on the sample of EDF cadets and conscripts? The main objective of this phase was to test how a handful of selected constructs/factors (decision-making style, unit cohesion and motivation) are for predicting the results of SJTDMs and whether these predictions are influenced by experiences acquired in the military sphere.

The sample used in this phase was the same as in phases 2 and 3 (see Table 2). The main instrument used for data gathering and the procedures were also the same as used in phases 2 and 3 (see more about sample, instrument development and quality improvement also in Article **III** and Article **IV**). The data gathered with the main instrument (SJTDM) and used in analyses had previously passed the process of quality improvement (see step 3 of phase 2) based on IRT analyses. However, during the data gathering (described in phase 2), there were additional questionnaires administered to the participants, which in addition to demographics included some instruments that were adapted to EDF needs and which at that time were part of a human resource management project (Allik & Talves, 2016). Those instruments were (see Article **V**):

- Unit Cohesion scale/measure that follows the principles of the standard model, and is based on the works ofAhronson & Cameron (2007), Griffith (2006), Ivey and Kline (2009), Shamir, Brainin, Zakay and Popper (2000) and was adopted for Estonian military context by Meerits (2012).
- Motivation at Work Scale (MAWS). The MAWS scale/measure (Gagné & Deci, 2005) was used to assess the influence that the type and degree of motivation has on personal performance.
- General Decision-making Style (GDMS). GDMS is a scale developed by Scott and Bruce (1995) for measuring commanders'/officers' leadership styles.

This additional data was used for analyses, which were concentrated in predicting SJTDM results. At first, it was tested how every selected measure predicted SJTDM results separately. Then, based on theory, a conceptual model (see Figure 3) was proposed and tested (see more about theoretical considerations on the basis of which the theoretical model was compiled in Article **V**). In the final model, covariance between expertise and secondary group cohesion was allowed in calculating estimates in SEM model (Article **V**).

At first, internal consistency (Cronbach alpha) of every used measure was checked by using IBM SPSS Statistics version 25. Acceptable reliability score was expected to be between .70 and .90 (Nunnally, 1978). Then, the pre-defined factor structure of each measure used was checked by confirmatory factor analyses (CFA). The purpose of this step was to make sure that all the instruments used in further analyses met the necessary criteria. After that, structural equation modelling (SEM) was used to verify the factors that predict the results of SJTDMs in the conceptual model (proposed in Article **V**). In the conceptual model, only decision-making measured by SJTDM is a latent variable, all independent variables are aggregated (Figure 3). The path coefficients, their statistical significance, and model fit indices were estimated. In the conceptual model, covariance between expertise and secondary group cohesion was allowed.

The measurement model was evaluated by normed chi-square (χ^2/df), the comparative fit index (CFI), the Tukey-Lewis index (TLI), the root mean square error of approximation (RMSEA) and the root mean square residual (SRMR). The following threshold values were applied: CFI > 0.90; TLI > 0.95; RMSEA < 0.08, χ^2/df< 5 (Byrne, 2016) and SRMR < 0.08 (Marsh, Hau & Grayson, 2005).

CFA and SEM analyses were performed using the software package IBM SPSS Amos 25 Graphics.

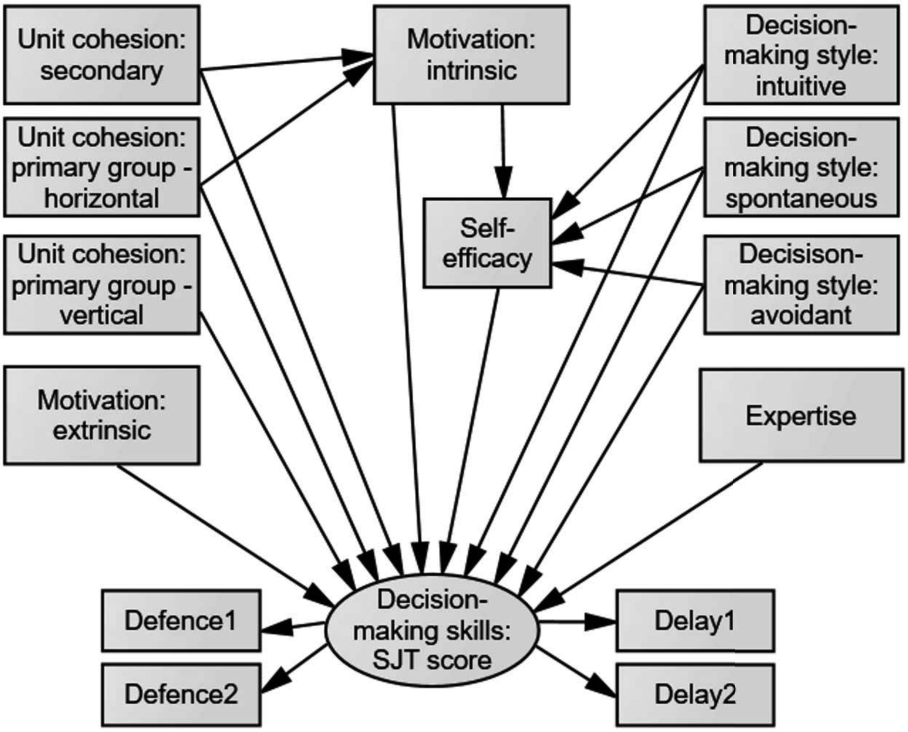

Figure 3. Conceptual model for predicting platoon level military commanders' performance on the SJTDM (Article **V**).

4. RESULTS

4.1. Phase 1: Preparation for instrument development

The ultimate goals of phase 1 were first to identify the basic knowledge for developing an instrument for assessing decision-making skills, and then to find, adapt or develop an instrument that can be used to assess the decision-making skills of low level military commanders in a battle-leading context. The first phase concentrated on preparation for instrument development and consisted of two clearly distinguishable steps. They were meant to answer research questions 1 and 2 of the doctoral thesis. The first step was meant to answer the first research question: what are the specific characteristics of a problem and problem solving in the military profession and how does it deviate from the general definition of a problem? The second step was meant to answer the second research question: what instruments have been used for measuring decision-making in battle (critical situations) and which instruments are best applicable and scientifically credible for measuring military commanders' decision-making skills?

4.1.1. Step 1. Specific characteristics of a problem and problem solving in military sphere

This first research question was addressed in Article **I** with the aims of specifying and describing the specifics of a problem and its characteristics in military sphere. The results indicated that in the military context, there are some specific characteristics of a problem and problem solving that should be highlighted in the context of the current doctoral thesis. In particular, it was concluded in Article **I** that in the military context, it is always necessary to consider the problem not only in the context of a practical military situation but also to stress the importance of a "bigger picture", which gives the problem a broader military background. In addition, it was stressed in Article **I** that while solving a problem (even if it is done under a command given), acting while carrying out the problem is very important in the military context. Another important aspect of problem solving in the military context that was brought out in Article **I** is that a problem still exists even if the problem solver cannot recognize it. One more difference highlighted in Article **I** was that in the military, the problem solver usually acts as the member (or commander) of some group (unit) of people and must cooperate with other units (higher, lower, neighbours). Thus, it must be also stressed as a specific characteristic that problem solvers operate in the context of a strict military subordination system, where one has to solve the problem as a result of an order given. This however also means that problems cannot be ignored even though they might not offer any cultural, intellectual or social value to the particular problem solver – they must be solved anyway and cannot be overlooked as the definitions typically suggest. Moreover, in the military context, an unsolved problem can escalate to a bigger problem for the problem solver or for someone else later.

Finally, the acceptance of losing human life as opposed to the unclear price of the goal makes problem solving in the military profession a bit different from the general problem solving process, as highlighted in Article **I**.

4.1.2. Step 2. Types of instruments/tests, selecting the instrument for current doctoral thesis

The second research question was addressed in a systematic literature review (Article **II**) with the objective of mapping how decision-making skills have been measured in military critical situations, and what specific instruments have been used to measure these.

The results indicated that only a few papers identified in the systematic literature review claimed to measure decision-making skills in similar circumstances to the interest posed in the current doctoral thesis (Article **II**). Yet the usefulness of even those papers was not obviously clear in the context of military battle leading. Despite the fact that in many of those papers, some sort of decision-making skills were measured, in many cases it was not used as primary measurement, but only as a secondary instrument for predicting purposes. Nevertheless, there were a handful of useful papers, based on which it was concluded that decision-making in critical, battle-like situations can be generally categorised into three types: (i) live performance; (ii) simulated performance; and (iii) tested performance (Article **II)**.

In the case of live performance and simulated performance, participants are typically placed into a high-stakes critical (real or simulated) environment and asked to solve certain tasks. Performance is then measured by an observer, usually an expert or by the simulator system itself. In the case of tested performance, participants are not asked to undertake a specific task in a certain (real or simulated) environment, but instead are presented with hypothetical scenarios and asked to solve some sort of test, where the questions raised are actually promted performance options. The participants are just asked to fill in the test by choosing the best one(s) from the proposed set of solutions (Article **II**).

It was therefore concluded that tested performance types of instruments are optimal for measuring military tactical-level decision-making skills (see all the considerations in Article **II**). The instrument that represents the tested performance type of test the best was found to be the situational judgement test. Thus, based on Article **II**, SJTDMs was the type of test selected for further development to be used as an instrument in the current doctoral thesis.

4.2. Phase 2: Instrument development, data collection, instrument quality control and improvement based on empirical data

Phase 2 was based on the results of previous phase, where it was decided that tested performance type of tests, namely the situational judgement test, need to be developed in the current doctoral thesis. This phase thus focused on answering the third research question: how to adapt (or develop) the instrument for EDF purposes and how good are its quality indicators? The objective was to describe in detail the process of instrument development, also report its quality indicators and improve the instrument if necessary.

4.2.1. Instrument development and data collection

The process of instrument development and data collection is described already in section 3.2. (as a part of methodology) of the current PhD thesis, in addition, the detailed process of construction, expert validation of the instrument and data collection were described in Articles **III** (in English) and **IV** (in Estonian).

As a result of the instrument development process, the research instrument SJTDM was developed and it consisted of three parts: 1) the general scenario, i.e. introduction of the general situation; 2) the initial (first) situation and the first decision-making task including answer options and an open response option; 3) the continuous situation and the second decision-making task including answer options and an open response option. Additionally, answer options were also a part of the instrument (see http://sisu.ut.ee/sjt/).

4.2.2. Instrument quality control and improvement

After conducting the tests (gathering data), one-parameter item response theory (1 PL IRT) analyses were undertaken in order to assess and improve the quality and reliability of the SJTDM (Article **III**). First, the internal consistency of the SJTDM test was assessed as a preparatory step, the initial results turned out to be rather low (.591) (see Table 3). After that, IRT analyses were conducted and selected quality parameters (see section 3.2.) were assessed.

Table 3. The internal consistency of SJTDMs before and after modfications based on the IRT analysis.

	n	N of items	Initial Cronbach's Alpha	Cronbach's Alpha after IRT analyses
SJTDM*	214	24	.59	.78

SJTDM – Situational Judgement Tests measuring Decision-Making Skills

The goodness of scoring key was assessed first. The results indicated that due to the similarity in the ability mean of some answer options and the small number of respondents, it was reasonable to use the 6-point correctness scale only for seven items. For all the other items a 2- to 5-point scale was more appropriate. The answers to these items were re-scored based on the results shown in Annex 1. In this process, it was also analysed if any revision is meaningful content-wise (Article **III**). For all the subsequent analyses in the framework of this doctoral thesis, new scores were used.

The goodness of SJTDM items in estimated discrimination the respondents was assessed secondly. The results show that most of the items are between of the suggested values of .5 to 2.0. (see Table 4) and thus appear to be very good for discriminating between respondents. There was only one item (k23) with a very low estimated discrimination score and thus outside of the suggested values. The item k23 seems to discriminate only between the respondents with very clear differences in decision-making skills (see more in Article **III**).

The goodness of item fit of the SJTDM test items was assessed thirdly. The results expressed as correlation coefficients (indication of how well the respondents' answers to each particular item correlate with their total test score) were usually above the recommended threshold .20 (see Table 4). Only in case of one item (v21) it was slightly below the suggested threshold. Taken together, this shows that most of the test items had a good fit in the SJTDM.

The variation of the difficulty measure of the SJTDM test items was assessed fourthly. The measure score of the test items (see Table 4) ranged from −1.36 (the simplest item) to .76 (the most difficult item) in the scale where the item difficulty measure was centred to zero. It appeared that nine items where simpler and 15 items more difficult than an average item, indicating that the SJTDM were quite well balanced. The average difficulty measure was 1.08, which shows that the SJTDM were rather simple for the respondents in the sample used.

The last step was to assess the internal consistency of the SJTDM based on revised (changed) score values (in accordance to Annex 1). After the changes made in scoring schema, the internal consistency was assessed once again. It appeared that the internal consistency of the test had improved significantly (from .591 to .780), which is between the acceptable values .70 and .90 (see Table 3). This indicates that the test scores are not dependent on the sample. It was therefore concluded that after the changes in scoring schema (result of IRT analyses), all of the SJTDM quality indicators were good enough to continue with further analyses in phases 3 and 4 of the current doctoral thesis. A detailed description of IRT analyses is available in Article **III**.

Table 4. 1PL IRT analysis results to describe 24 test items (n = 214) (Article **III**).

Item	Measure*	Correlation**	Estimated discrimination***
k11	.39	.42	.82
k12	–.08	.35	.99
k13	.14	.38	.93
k14	.39	.42	1.04
k15	.06	.37	1.11
k16	–.12	.34	1.04
k21	.36	.42	.99
k22	.32	.41	1.29
k23	.76	.48	.05
k24	–.05	.35	1.19
k25	.10	.38	1.15
k26	–.71	.26	.98
v11	.28	.40	.81
v12	.41	.42	1.49
v13	.45	.43	.92
v14	–.21	.33	.97
v15	.35	.42	1.12
v16	.16	.39	.87
v21	–1.35	.19	1.01
v22	–.99	.23	1.00
v23	.15	.38	1.06
v24	.06	.37	.99
v25	–.24	.33	.97
v26	–.65	.27	.98

*measure of item difficulty

**the correlation between each item score and respondent ability score (total score of the test), which is calculated based on the model, not based on the total score of the items

*** how well each item is estimated to discriminate the persons if a 2PL IRT model would be used instead of the 1PL IRT model; a result from .5 to 2.0 is considered as good (in one parameter IRT, discrimination index gives an estimate of what would be the discrimination of the item if a two-parameter model was used in which discrimination is treated as a free parameter that can be assigned any value)

4.3. Phase 3: Describing the results of the test based on different groups

Phase 3 is based on the results of Phase 2, where the quality indicators of developed SJTDM were checked and the test was improved by introducing new scoring schema based on IRT analyses. The SJTDM were thus considered suitable for further analyses in phases 3 and 4.

Phase 3 focused on answering the fourth research question: how accurately does the instrument distinguish novices from experts/professionals in a simulated platoon leader battle scenario in the example of EDF cadets (advanced) and conscripts (novices)? The objective was to describe the differences in SJTDM results, the number of problems identified in situations and time spent to solve the tests. All of that was described based on whole sample and also based on different groups, which were created based on test takers' experience (more and less advanced) and the way, in which the situations were presented (paper and video). The fourth research question was addressed in detail in Article **IV.**

In looking at the combined SJTDM results based on independent groups, it was expected that more experienced test takers (cadets) would be better in solving SJTDM than less experienced test takers (conscripts). In addition, due to the differences in the information available, it was expected that test takers in the paper group (allowed to read the information about the situation on paper) would get better results than test takers in the video group (received the same introduction while watching the video). The results demonstrate (Table 5) that indeed cadets (average score 121.3 points) solved the SJTDM better than conscripts (112.9 points), and the difference was also statistically significant ($t = -6.257$, $p < 0.001$). Thus, previous experience appears to be necessary for achieving better results. This confirmed that the test objectively shows differences in decision-making skills. However, it appeared that there was no statistically significant difference in the SJTDM result (Table 5) based on the way in which the situations were presented to the test takers (paper and video). Thus, the way the situations were presented did not affect the test results. This might also be considered a valuable outcome, showing that in using the test in the future there are always two options to choose from.

In comparing the combined test results to initial situations (simpler) and continuous follow-up situations (more difficult), it was expected that solutions to initial situations would be better than in case of continuous situations. However, the results show (Table 5) that the solutions to continuous situations (60.7) were better than the solutions to an initial situation (57.4). The difference is statistically significant ($Z = -7.544$; $p < 0.001$).

In looking at the number of problems that test takers were able to identify in the situations, it was expected that more experienced test takers (cadets) would be able to detect more problems in situations than less experienced (conscripts). In addition, it was expected that test takers in the paper group (allowed to read the information about the situation from the paper) could identify more problems in situations than test takers in the video group (received the same introduction while watching the video). The results show (Table 5) that there were no differences between more experienced (cadets) and the less experienced (conscripts) test takers. However, in the case of the different groups based on situation presentation (paper and video), the results show that the test takers who belonged to the paper group were able to identify slightly more problems in situations than the test takers in the video group (in an average 11.5 and 10.41, respectively). The results however are not statistically significant ($t = –1.952$; $p = .052$), although the direction of the results is as expected.

In comparing the number of problems identified in the initial situations (simpler) and continuous situations (more difficult), it was expected that more problems would be identified in the case of continuous situations. The results (Table 5) show that indeed more problems were identified in continuous situations than in initial situations (an average 5.9 and 5, respectively). The results were statistically significant ($Z = –6.102$; $p < 0.001$).

Table 5. Results of independent samples analyses (table compiled based on the results presented in Article **IV**).

Independent samples analyses (t-test): **cadets and conscripts.**				
	Cadets	**Conscripts**	**t**	**p**
Test results (combined)	121.3 points	112.9 points	–6.25	.000
Number of problems identified (combined)	11.1	10.9	–.405	.686
Time (combined)	1953 seconds	1858 seconds	–.367	.173
Independent sample analyses (t-test): **paper and video**				
	Paper	**Video**	**t**	**p**
Test results (combined)	118.7 points	117.6 points	–.747	.456
Number of problems identified (combined)	11.5	10.4	–1.952	.052
Time (combined)	2113 seconds	1703 seconds	–6.714	.000

Looking at the time it took test takers to solve the SJTDM, it was expected that more experienced test takers (cadets) could solve the tests faster than less experienced ones. It was also expected that test takers from the video group (who received the introduction from watching a video) were able to solve the tests faster than test takers in the paper group (allowed to read the information about the situation off paper). The results (Table 5) showed that there was no statistically significant difference in the time it took to solve the tests between less experienced test takers (conscripts) and more experienced test takers (cadets).

In the case of different groups based on situation presentation (paper and video), the results (Table 5) show that the test takers who belonged to video group were indeed able to solve the tests faster than test takers in the paper group (an average 1703 and 2112 seconds, respectively). The results were statistically significant ($t = -6.714$; $p < 0.001$).

Comparing the time it took test takers to solve the initial situation (simpler) and continuous follow-up situation (more difficult), it was expected that solving more complex (continuous) situations would also take more time. However, the results (Table 6) showed just opposite – test takers were actually able to solve continuous situations faster than initial situations (in an average 849 and 1068 seconds, respectively). The results are statistically significant ($Z = -11.365$; $p < 0,001$). This might be explained by the fact that in the continuous situation, knowledge from the initial one was used.

Table 6. Results of dependent samples analyses (table compiled based on the results of Article **IV**).

Dependent samples analyses (Wilcox on Signed ranks test): **initial and continuous situations**				
Variable	**Initial**	**Continuous**	**Z**	**p**
Test results (combined)	57.4 points	60.7 points	–7.544	<.001
Time (combined)	1068 seconds	849 seconds	–11.365	<.001
Number of problems identified (combined)	5.0	5.9	–6.102	<.001

4.4. Phase 4: Factors predicting results of SJT

Phase 4 is based on the results of phase 2, where the quality indicators of the developed instrument were checked and the instrument was improved by introducing new scoring schema based on IRT analyses. The tests were thus considered suitable for further analyses in phases 3 and 4. Phase 4 focused on answering the fifth research question which was: based on the example of EDF cadets and conscripts, what attributes predict military commanders' decision-making skills at the platoon leadership level? The objective was to identify what are the factors that predict better test results. The fifth research question was addressed in detail in Article **V.**

At first, the measures of service motivation (MAWS), general decision-making style (GDMS) and unit cohesion were one by one tested in SEM models predicting test results (Article **V**). Goodness of fit data of each used measure was tested first by CFA analyses, the resulted data is shown in Table 7.

In the case of the SEM model with Unit Cohesion, it was expected that positive primary and secondary unit cohesion would be good predictors of SJTDM because support from superior officers and peers has previously been argued to have a positive influence on subordinates. The model gave good fit values (Model 1 in Table 8) for predicting SJTDM. However, it appeared that only two factors out of four (secondary group cohesion and primary group vertical cohesion) were statistically significant for predicting SJTDM (Article **V**). Surprisingly, the parameter estimates indicate that commanders' influence predicted SJTDM results negatively. It also appeared that the influence on perceived institutional cohesion predicted SJTDM results positively (Article **V**).

Table 7. Goodness of fit data of the measures used in the doctoral thesis (based on CFA models) (Article **V**).

	χ^2	DF	χ^2/DF	CFI	TLI	RMSEA	SRMR
SJTDM (4 factors)	0.786	2	0.39	1	1	0.00	0.012
Unit Cohesion (4 factors)	208.13	98	2.12	0.93	0.91	0.073	0.076
MAWS (2 factors)	11.80	8	1.47	0.99	0.98	0.047	0.042
GDMS (5 factors)	241.57	160	1.51	0.90	0.88	0.049	0.074

Note: χ^2 – chi-square statistics, DF – degrees of freedom, CFI – comparative fit index, TLI – Tukey-Lewis index, RMSEA – root mean square error of approximation. **p < .001, *< .05, SRMR – standardized root mean square residual.

In the case of the SEM model with MAWS, it was expected that internal motivation would predict better SJTDM results. The two-factor model (Model 2 in Table 8) gave good fit values; the results showed that external motivation predicted lower SJTDM results and internal motivation predicted higher SJTDM results (Article **V**).

In the case of the SEM with GDMS, it was expected that intuitive decision-making style would predict higher SJDM results; and avoidant and dependent decision-making styles, lower SJTDM results. The model fit was good (Model 3 in Table 8), however, none of the factors appeared to be statistically significant predictors of decision-making skills. The overall trend however was as expected: intuitive decision-making style predicted higher, avoidant and dependent decision-making styles predicted lower SJTDM results (Article **V**).

Table 8. Fit indices of the Structural Equation Models (Article **V**).

Models	χ^2	**DF**	χ^2**/DF**	**CFI**	**TLI**	**RMSEA**	**SRMR**
Model 1 (Unit Cohesion)	272.25	160	1.70**	0.93	0.92	0.057	0.069
Model 2 (MAWS)	35.58	32	1.11	0.99	0.99	0.023	0.041
Model 3 (GDMS)	322.85	237	1.36*	0.91	0.89	0.041	0.070
Model 4 (conceptual model)	84.55	60	1.41*	0.96	0.95	0.044	0.056

Note: χ^2 – chi-square statistics, DF – degrees of freedom, CFI – comparative fit index, TLI – Tukey-Lewis index, RMSEA – root mean square error of approximation. **p < .001, *< .05, SRMR – standardized root mean square residual.

Secondly, the conceptual model was tested, which besides the direct effects considered the indirect effects and associations between exogenous variables (Figure 3). In the conceptual model, covariance between expertise and secondary group cohesion was allowed during the calculation. The model fit was good (Model 4, Table 8), indicating that the conceptual model quite adequately fits the data. All the paths and standardized regression weights of each independent variable are shown in Figure 4. The path coefficients for statistical (p-values) and practical significance (standardized regression weights, β) and correlations are also shown in Table 9.

Extrinsic and intrinsic motivation were expected to have a direct effect on SJTDM. Results show that higher level of extrinsic motivation predicts decrease of the results of SJTDM ($\beta = -0.16$, $p < 0.05$). However, intrinsic motivation has no statistically significant direct effect on SJTDM results ($\beta = 0.08$, $p = 0.50$) (Article **V**).

All types of unit cohesion were expected to have a direct effect on SJTDM. The results show that out of primary group cohesion, only vertical group cohesion has a statistically significant negative ($\beta = -0.18$, $p < 0.05$) direct effect on SJTDM. Horizontal primary group cohesion and secondary group cohesion have very small but statistically not significant direct effect on SJTDM (Article **V**).

Intrinsic motivation was expected to mediate the effects of secondary and primary group horizontal cohesion and also influence the self-efficacy. The results showed that indeed secondary group cohesion ($\beta = 0.78$, $p < 0.001$) and horizontal primary group cohesion ($\beta = 0.12$, $p < 0.01$) have a statistically significant effect on intrinsic motivation. However, the total indirect (mediated) effect of secondary group cohesion on SJTDM appeared to be statistically not significant. The indirect (mediated) effect of horizontal primary group cohesion on SJTDM appeared to be positive but very modest ($\beta = 0.01$). Thus the results indicate that high cohesiveness with superiors might decrease the results of SJTDM. The results also indicate that there is a strong positive association between secondary group cohesion and intrinsic motivation but neither has a significant effect on SJTDM. In addition, the primary group horizontal cohesion effect on SJTDM might be higher in the case of higher intrinsic motivation (Article **V**).

Intuitive, spontaneous and avoidant decision-making styles were expected to have a direct effect on SJTDM. Self efficacy was also expected to have direct effect on SJDM, in addition, it was expected to mediate the effects of intuitive, spontaneous and avoidant decision-making styles and intrinsic motivation. The results showed that none of the decision-making styles predicted SJTDM in a statistically significant manner. However, self-efficacy has statistically significant ($p = 0.09$) but weak direct effect ($\beta = 0.13$) on SJTDM. Out of all decision-making styles used in the conceptual model, only the spontaneous decision-making style ($\beta = 0.19$) has statistically significant ($p < 0.01$) positive effect on self-efficacy. The intrinsic motivation also has no statistically significant effect on self-efficacy. It seems thus that self-efficacy has a very small positive mediating effect (0.03) on SJTDM only in the case of spontaneous decision-making style (Article **V**).

Table 9. Estimates of conceptual model (standardized coefficients) (Article **V**).

Regression Weights			**Estimate**	**P**
Unit cohesion: secondary	→	Intrinsic motivation	0.781	***
Unit cohesion: primary horizontal	→	Intrinsic motivation	0.123	<0.01
Decision-making style: avoidant	→	Self-efficacy	–0.074	.274
Decision-making style: spontaneous	→	Self-efficacy	0.193	<.01
Decision-making style: intuitive	→	Self-efficacy	0.074	.309
Intrinsic motivation	→	Self-efficacy	0.094	.155
Self-efficacy	→	Decision-making skills (SJTDM)	0.135	.091
Unit cohesion: secondary	→	Decision-making skills (SJTDM)	0.183	.216
Decision-making style: spontaneous	→	Decision-making skills (SJTDM)	–0.077	.339
Intrinsic motivation	→	Decision-making skills (SJTDM)	0.077	.530
Expertise	→	Decision-making skills (SJTDM)	0.279	<.05
Decision-making style: avoidant	→	Decision-making skills (SJTDM)	–0.091	.227
Unit cohesion: primary vertical	→	Decision-making skills (SJTDM)	–0.184	.048
Unit cohesion: primary horizontal	→	Decision-making skills (SJTDM)	–0.074	.351
Extrinsic motivation	→	Decision-making skills (SJTDM)	–0.167	.064
Decision-making style: intuitive	→	Decision-making skills (SJTDM)	–0.021	.789
Decision-making skills (SJTDM)	→	Defence1	0.713	***
Decision-making skills (SJTDM)	→	Defence2	0.397	***
Decision-making skills (SJTDM)	→	Delay1	0.804	***
Decision-making skills (SJTDM)	→	Delay2	0.309	***
Correlations				
Unit cohesion: secondary	↔	Extrinsic motivation	0.449	***
Extrinsic motivation	↔	Expertise	0.313	***
Unit cohesion: secondary	↔	Unit cohesion: primary vertical	0.312	***
Unit cohesion: secondary	↔	Expertise	0.647	***
Unit cohesion: secondary	↔	Unit cohesion: primary horizontal	0.175	<.001
Decision-making style: spontaneous	↔	Decision-making style: intuitive	0.394	***
Decision-making style: intuitive	↔	Decision-making style: avoidant	0.192	<.01
Decision-making style: spontaneous	↔	Decision-making style: avoidant	0.096	.162
Extrinsic motivation	↔	Unit cohesion: primary vertical	0.299	***
Extrinsic motivation	↔	Unit cohesion: primary horizontal	0.168	<.05
Unit cohesion: primary vertical	↔	Unit cohesion: primary horizontal	0.351	***

Expertise was expected to have a direct effect on SJTDM. The results clearly indicate that expertise has a strong (β = 0.28) and statistically significant ($p < 0.05$) positive influence on SJTDM results. It supports the results of phase 3, which show that cadets (more experienced) performed better than conscripts (less experienced) (Article **V**).

The conceptual model explained 26% of the variance of the latent variable decision-making skills measured with SJTDM. In the final model, covariance between expertise and secondary group cohesion was allowed during the calculation. The strong and statistically significant correlation ($r = 0.65$, $p < 0.001$) revealed that more experienced test takers (cadets) were also more committed to the EDF and are more internalized (have taken over the values and aims of the EDF) than conscripts (Article **V**).

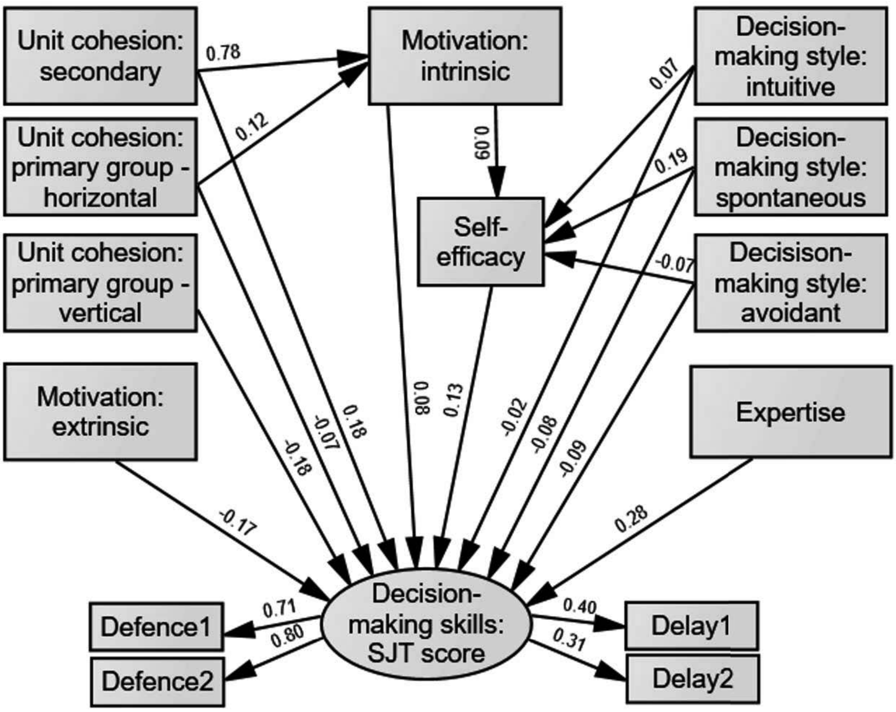

Figure 4. Conceptual model with standardized path coefficients (decision-making skills is latent variable, all others are aggregated variables) (Article **V**).

5. DISCUSSION

5.1. Knowledge needed to develop a test for assessing decision-making skills in military context

Firstly, as a point of departure for this study, the nature of a problem and problem solving in the military sphere was of interest. Some of the findings were merit a more detailed discussion before going on to the next step of phase 1.

For one thing, a problem was typically defined only as an unknown entity between the current state and goal state, for which the problem solver has to find a value. The unknown in this formula is the difference between the current state and the goal state in a situation (Jonassen, 2000). Sometimes it was briefly mentioned that a problem requires a practical and motivating environment (Meacham & Emont, 1989) and situation (Jonassen, 2000). Some authors have described problems through the need to achieve some goals and in this case a problem occurs if the problem solver has to achieve an objective (Mayer & Wittrock, 1996). The value for the problem solver has been argued by some other authors as well; for example Arlin (1998) adds, that a problem is not a real problem if there is no "felt need", e.g. amotivation to solve it (Arlin, 1989). So, obviously a problem must offer some social, cultural, or intellectual value for the problem solver (Jonassen, 2000). Some authors stress that an issue that comes up in the case of a problem is that It is not exactly known how to reach the goal state and when doing so, problem solvers have to act (Robertson, 2001).

The current study shows that the definition of a problem is well suited to the military context, but there are some differences. Although some authors briefly mentioned the need for a practical and motivating environment (Meacham & Emont, 1989) as well as situation (Jonassen, 2000), a problem was still typically defined only as unknown entity in between the current state and goal state. But in the military context, it is always necessary to consider the problem not only in the context of practical military situation, but also stress the importance of the "bigger picture", which gives the broader military background to the problem. In addition, one difference is that in the military, the problem solver almost always acts as the member (or commander) of some group (unit) of people and has to cooperate with other units (higher, lower, neighbours). Thus problem solvers operate in the context of strict military subordination system, which means that problems are given to solve as orders.

Robertson (2001) stressed the importance of acting while solving a problem and this seems to be very important especially in the military context (thinking in the absence of acting might not help you with the enemy). Here acting is considered as something that happens as the result of an order in some practical situation, which also happens in the context of bigger picture. If this is the case, the problem still exists even if the problem solver cannot recognize it as described by Arlin (1989) and Jonassen, 2000). Additionally, problems cannot be ignored even though they might not offer any cultural, intellectual or social value for the solver (as for example killing), but they have to be solved anyway and cannot be ignored as the definitions typically suggest (Arlin, 1989; Jonassen, 2000; Nitko, 2001). The acceptance of losing human life and the unclear price of the goal (compared to the value of human lives lost) makes problem solving in the military profession especially different from the general definition of a problem (and different from other professions as well).

To conclude, a problem in the military profession can be defined as an unknown entity in between the current state and goal state in some practical situation, which at the same time is a part of the "bigger picture". The biggest difference in the military profession (compared to the general problem definition) seems to be the acceptance of losing lives and the price of the reached goal compared to human lives lost while achieving it. In addition:

- The problem solver does not know exactly how to reach the goal and while solving the problem, there is a need to apply the solution in some practical situation or – at least to some extent – to act (in addition to thinking) while solving the problem.
- The problem still exists if the problem solver fails to recognize the unknown entity between the current state and the goal state (in this situation, the problem might even escalate and become a problem for someone else.
- It is not possible to ignore the problem in the military profession; therefore, it is not an option of problem solver.

Based on the findings of the step 1 of phase 1, it can be concluded that most common characteristics of a problem in the military context are the constantly changing nature of the situation, where problem solving occurs, and the need for the problem solver to deal with such a situation. In his classification of the problems, Jonassen (2000) lists problems based on the criteria that can be designed for solving them in a learning context. As for the type of problem one is likely to face in the military context, Jonassen (2000) terms these strategic performance problems. This type of problem is characterized by the need to make decisions in a real-time complex and changing situation while maintaining situation awareness (Jonassen, 2000). The strategic performance type of problems were thus selected to be the ones used in the next phases of this doctoral thesis (Article **I**) – especially when it comes to developing the SJTDM.

The second step of phase 1 was a systematic literature review undertaken to find out whether the type of testing tool described in the previous step of phase 1 already exists or not. If the findings indicate the existence of such a test, the interest turns to reusing or adopting it. If such a test does not exist, the interest turns towards identifying at least some existing test types that can be used as a basis for developing new tests for this doctoral thesis. If there is nothing, the interest turns to making the testing tool from scratch based on the gathered test samples. The criteria for inclusion and exclusion of articles, and also the criteria for evaluating selected articles, are provided in section 3 and in detail in Article **II**.

It appeared that the majority of the identified papers during literature review ended up with measuring different types of constructs, which were only supported by some sort of decision-making measures (Frederico, 1997; Nesbitt, Kennedy, Alt & Fricker, 2015). However, decision-making, not to mention decision-making in the military battle leading context, was rarely the main focus of research (Article **II**). In cases where decision-making skills were measured at least close to military context, the sample of the studies has usually been university students (Chancey & Bliss, 2012; Hale, Stanney & Malone, 2009; Keebler, Jentsch & Schuster, 2014; Saus, Johsen, Eid & Thayer, 2012; Vogel-Walcutt, Carper, Bowers & Nicholson, 2010). In those cases, the conducted tests were typically also unrealistic for the military context. For example, searching for a briefcase alone in an enemy-guarded territory is not a realistic task for a normal military unit commander (Chancey & Bliss, 2012; Hale et al, 2009). Nevertheless, many of the identified papers were suitable for the focus of our research. The list of suitable articles is provided in Article **II**.

Further, based on the suitable studies, three categories of the test types were proposed on the basis of the type of decision-making performance measured in the study: (i) live performance, (ii) simulated performance, and (iii) tested performance (Article **II**). In the case of live performance and simulated performance, participants are placed in a real high-stakes critical (or simulated) environment and asked to solve certain tasks. Their performance is then observed and evaluated by an expert in the case of live performance and by the simulator system itself in the case of simulated performance (Article **II**). However, due to the amount of resources required to prepare and conduct the tests (or procure the simulator system), these types of tests were not considered suitable for the current doctoral thesis. More detailed argumentation and a list of the tests can be found in Article **II**.

Since the evaluation criteria were about being related to the military sphere in one way or another, cost-effectiveness and simplicity in developing and conducting the tests, it was concluded that the most optimal test type in the Estonian context would be tested performance (Article **II**). It differs from the previous two because in such tests, participants did not actually have to perform in a realistic environment. Instead they were presented with hypothetical what-if scenarios, which they had to solve either by filling in their own solutions or choose the best one(s) from a pre-prepared set of solutions or order these solutions by their effectiveness. Examples of such tests discovered by literature review were conducted by Shadrick and Lussier (2004), Nesbitt et al (2015), Connelly et al (2000), Horvath et al (1996), Hauenstein, Findlay and McDonald (2010) and Frederico (1997). Shadrik and Lussier (2004) assessed the critical-incident-determining skills of US Army Captain Career Course students. Nesbitt et al (2015) explored the possibility of adapting Iowa Gambling Task to the military domain. Connelly et al (2000) tested the Leader Capabilities model using a sample of US Army officers in problem-solving during a leadership course. Horvath et al (1996) used the theory of tacit knowledge to test military commanders in different scenarios. Hauenstein, Findlay and McDonald (2010) used the participants of the Equal Opportunity Advisors course as a sample in order to fix scoring keys for training SJTs. Frederico (1997) used a sample of naval officers to test whether metacognitive models of the abstract components of situation assessment were correlated with performance on concrete experimental tasks that necessitate situation assessment.

Of the best suited studies in the literature review (Article **II**), it was found that most of them consisted of the scenarios and situations given to test takers together with pre-prepared answer options to a given situation. In the case of answer options, test takers were typically asked to choose the best option or to rank the options by solution quality. The exact test type that came across frequently was SJT. Thus, based on the findings of phase 2, it was decided to use it in this doctoral thesis as well. The findings indicated that no particular test was up to the task of measuring military commanders' decision-making skills in a battle leading environment. It was thus concluded that there was a need to develop such test from scratch (Article **II**).

5.2. Quality of the compiled test for measuring military commanders' decision-making skills

As was concluded in phase 1, the development of an instrument called SJTDM had to be undertaken in phase 2. Before starting development, it had to be considered that there is a widespread understanding, that SJT results can reflect a variety of multiple different constructs or composite competencies (Chan & Schmitt, 2005) such as applied social skills, heterogeneous composites, leadership, basic personality tendencies, job knowledge and skills, and teamwork skills (Christian, Edwards & Bradley, 2010). However, from the point of view of the current doctoral thesis, the most important is that SJT responses can also be expected to be a function of generic and domain-specific job knowledge gained through experience or formal education (Ployhart & Weekley, 2006). Consequently, it would be possible to develop SJTs that mostly reflect the learning outcomes of some specific military topic, in our case decision-making in a battle context. This was kept in mind in developing the SJTDM.

While developing the general scenario, specific situations and especially answer options, it was attempted to compile all of that such that the SJTDM above all would reflect situations' domain specific practical knowledge and experience. Experts were used for validating the general scenario, situations and response options as recommended by Bergman et al (2006).

After that, it was initially decided to test the SJTDM developed, which are based on empirical data. For this, IRT analysis was chosen, because it is based on Bayesian statistics, which is more robust in such analyses than typical classical psychometric theory analyses, which are based on factor analyses; in the case of items with different difficulty, this could result in biased results (Article **III**). The results indicated that the quality of the SJTDM was already good, but based on the results of IRT analysis, it was possible to even improve it for further analysis. It means that SJTDM can be used for assessing low level military commanders' decision-making skills in a simulated battle leading environment. In addition, the procedure of developing and improving SJTDM based on empirical data can be used if there is a need for additional tests. Together with the fact that SJTDM is quite easy to develop, the implication is that based on our findings, the EDF now have a reliable tool for measuring their reserve officers' battle leading skills in a cost-effective manner.

5.3. Decision-making skills of the military commanders

When the quality of the SJTDM was checked, the next step was to find out how well our test takers performed on the SJTDM. This was done by comparing the results based on the whole sample, and also different groups. In this phase, most of the gathered data was used in analyses. In addition to test results, the data also included the number of problems identified in situations and test taking time. The detailed results are presented in Article **IV**.

Better trained and more experienced test takers (cadets) were expected to get a better result in SJTDM, solve the tests faster, and identify more problems in the presented situations. It has been discovered in many studies that more experienced officers are able to act faster and more effectively because they comprehend the situation better than those with less experience (Fuglseth & Grønhaug, 1995; Glaser, 1985; Rasmussen, 1983; Cohen-Hatton & Honey, 2015; Foldes et al, 2010; Grier, 2012; Norman, 2006; Sadler-Smith & Shefy, 2004; Vowell, 2004). In the same vein of these results, cadets clearly outperformed conscripts, confirming that more experienced test takers get better results and the developed test is suitable for recognizing these differences. It can therefore be concluded that military education/training and acquired experience resulted in better SJTDM results. However, in the case of time and problems identified, the results of the current study do not confirm the results of other authors. It means that in our study there was no difference between more (cadets) and less experienced (conscripts) test takers.

Encouraged by previous research, which indicates that there could be differences in SJTDM results based on how the scenario is presented to test takers (Richman-Hirsch et al, 2000; Tan, Tse & Chung, 2010; Chan & Schmitt, 1997; Whetzel & McDaniel, 2009), it was expected in this study that these differences would appear in the sample of this study as well. Especially test takers who received the presentation of the situation on the paper and were expected to get better results than those who received the presentation as a video. However, the results indicated no differences, which however seems to be in line with the results in the field of medicine education, according to the study by Lievens and Sackett (2006). In the current study, the reason could be that the SJTDM were not difficult enough for the sample for such differences to appear. However, the video group was able to solve the tests more quickly, probably due to the fact that they had no papers to browse while solving the tests and thus were able to work faster.

In Article **IV**, the differences between solutions to the initial situation and continuous situations were analysed and discussed, but since these findings are not in the focus of the current doctoral thesis, they remain to be explored from Article **IV**. The main findings of phase 3 indicate that the SJTDM is a reliable test to distinguish between better and worse test takers. Thus it can be confirmed that the SJTDM developed in this doctoral thesis are well suited for evaluating test takers with different experience (e.g. the ones with better training and experience get better results).

5.4. Predictors of military commanders' decision-making skills

In phase 4 the objective was to identify the factors that predict better SJTDM results. It focused on answering a fifth research question: based on the example of EDF cadets and conscripts, which attributes predict military commanders' decision-making on a platoon leadership level? Interest was paid to how much other factors influenced the SJTDM developed keeping in mind educational (training) context of EDF. Regardless of the interest towards predicting factors however, in such a context, it would be useful if the results of the SJTDM reflected mostly domain specific practical knowledge gained through training and previous experience in the military sphere. In such a case, training staff would be the best (without much influence other than the level of knowledge of test takers) informed of trainees' training level. Therefore, it would help to make adequate training related decisions.

Ployhart and Weekley (2006), for example, emphasize that instead of different constructs, SJT responses can be expected to be a function of generic and domain-specific job knowledge gained through experience or formal education. This idea of SJTs measuring only one construct seems to be supported by Johnson and Oswald (2010), who argue that analyses of SJTs usually result in a single general situational judgment construct. However, it has been noted that the content of SJTs can be developed in a way that test taking performance (i.e. test scores) reflect a combination of multiple constructs or composite competencies (Chan & Schmitt, 2005). Nevertheless, it seems that SJTs can be developed in a way that above all they reflect domain specific practical knowledge. In the case of this doctoral thesis, SJTs required very specific knowledge about military tactics and decision-making for getting better results.

In phase 3, it was already confirmed that the results of developed SJTDM depend on the level of training and thus also on previous experience, and discriminates beginners from more experienced test takers. Therefore, the further objective was to find out how much some other factors (other than the level of training and experience) influence the results of the SJTDM. After experimenting with constructs one by one, an attempt was made to combine several of the factors, including experience, into the model, which was also tested.

More specifically, in phase 4, attention was paid to the decision-making style, unit cohesion, motivation, self-efficacy and experience, as factors predicting the results of the SJTDM. These measures were selected from the different aspects of the conscripts' profile (as a part of the human resource project carried out in EDF since 2015) (Allik & Talves, 2016) based on the theory presented in Article **V**. After checking the effects of selected constructs one by one (each scale individually), different factors and variables were consolidated into the conceptual model (Figure 3) and tested (see Article **V**). Here, only the most important outcomes are discussed.

Based on the concepts proposed by Bartone and Kirkland (1991), Ryan, Deci & Edward (2000), Siebold, (2007, 2011), Fuglseth and Grønhaug (1995), Glaser (1985), Gagné and Deci (2005), Stajkovic and Luthans (1998), it was expected that perceived unit cohesion, self-efficacy, motivation and decision-making styles would directly influence military commanders' performance on solving military domain specific SJTDM. In addition to a direct effect, intrinsic motivation was expected to shape the association between horizontal primary group cohesion and secondary group cohesion as well as influence the effect of self-efficacy as proposed by Buch et al, (2016).

Despite the fact that Deci and Ryan (2000) have highlighted the importance of intrinsic motivation and autonomous regulation of extrinsic motivation in leading to positive psychological and performance outcomes, this turned out not to be the case in this study. As intrinsic motivation had no significant impact on the results of the SJTDM, extrinsic motivation appears to have a negative influence towards the SJTDM. Since perceived autonomy is argued to be vital in the case of motivation, the reasons behind this outcome could be that the level of autonomy is quite low in conscript training, and actually in overall daily schedule as well (Article **V**). However, the reason could be also explained by the findings of Kusukar et al (2012) who proposed that autonomous motivation affects academic performance positively not directly, but through a deep strategy towards learning and higher study effort. Learning strategies were not part of the conceptual model in the current study and thus the concept proposed by Kusukar et al (2012) remain to be discovered in future studies.

Self-efficacy had a very weak effect on the results of the SJTDM, but there was no significant mediating effect of self-efficacy on SJTDM results as proposed by Buch, Säfvenbom and Boe (2016), who showed that the impact of self-efficacy on perceived military performance depends on the level of intrinsic motivation. The reason might be explained by the results of Stajkovic and Luthans (1998), who proposed that the effect of self-efficacy on task performance is mediated by the complexity of the task. Looking at the cadets and conscripts separately, it appeared that the impact of self-efficacy was indeed more salient in the case of cadets to whom the task was less complex. This outcome fits in well with the findings of Stajkovic and Luthans (1998) and seems to prove that perceived self-efficacy without experience is simply not sufficient in the case of complex tasks.

Unit cohesion, especially primary group cohesion, is believed to be of critical importance in group and individual performance (Beal, Cohen, Burke & McLendon, 2003; Gully, Dennis & Whitney, 2012; Mullen & Copper, 1994; Oliver, Harman, Hoover, Hayer & Pandhi, 1999; Shils & Janowitz, 1948; Salo, 2006; Siebold, 2006; Jacobs, 1991; Mael & Alderks, 1993). This might apply to secondary group cohesion as well, because military cultures arguably provide a pre-existing cohesive background (Siebold, 2011). It turned out that the effect of primary group horizontal cohesion was not statistically significant, and the effect of primary group vertical cohesion turned out to be negative. Both results were unexpected, but seem to be in line with Pawiński's (2018) arguments that primary group cohesion could lead to negative consequences, such as the demotivated behaviour of soldiers. However, the explanation could also be related to something as simple as the fact that SJTDM were conducted in classroom settings, where the relations among group members were not as important as they could be in a battle situation. Moreover, these unexpected findings can perhaps be explained by the fact that test takers (conscripts and cadets) were in the middle of their military training and had recently been under different commanders and peers in different training courses. In line with primary group cohesion, the influence of secondary group cohesion appeared not to be a significant predictor of SJTDM; however, it appeared to be correlated with intrinsic motivation, which is in line with Siebold's (2011) remark about pre-existing cohesiveness in military culture. This suggests that commitment to the EDF and internalization of its organizational values promotes intrinsic motivation and thus indirectly has only a small positive predictive effect on the SJTDM.

In the case of decision-making styles, the views of Thunholm (2004, 2009) and Sadler-Smith and Shefy (2004) were followed, suggesting that intuitive and spontaneous decision-making styles could predict better results in the SJTDM and that avoidant style could predict lower results in the SJTDM. However, none of the decision-making styles appeared to have a significant influence on the SJTDM. The reason could be that in the case of solving tests in a stress free classroom environment, decision-making styles are simply not as important as they would be on a real battlefield. Nevertheless, the overall trend was observable and it coincided with Thunholm (2004, 2009) and Sadler-Smith & Shefy (2004).

Finally, following Fuglseth & Grønhaug (1995), Glaser (1985), Rasmussen (1983), Cohen-Hatton and Honey (2015), Grier (2012), Norman (2006), Vowell (2004), Sadler-Smith and Shefy (2004) recommending that expertise (the level of military education/training and acquired experience) result in better SJTDM results. In phase 3 of this doctoral thesis, it was already confirmed that more experienced test takers were better in solving SJTDM than less experienced ones (Article **III** and **IV**). The conceptual model resulted in exactly the same findings confirming that experience predicts SJTDM very well.

Looking at the conceptual model as a whole, it appeared that the model had indicators that were a good fit, and 26% of the variance of the latent variable (decision-making skills) measured by the SJTDM was explained. However, taken together, it appeared that, despite quite a solid theoretical foundation, good fit data and explanation of the variance of the latent variable (Article **V**), only a handful of factors other than what was already known from phase 3 (experience) had a small influence on the latent variable (results of the SJTDM in our conceptual model). In light of this doctoral thesis, this finding is however good, because the SJTDM developed and used in the study attempted to compile the measuring of very specific learning outcomes. Therefore, the result should not reflect anything other than respondents' generic and domain-specific job knowledge gained through experience or formal education, which is good to keep in mind. The findings thus seem to align with the ideas proposed by Ployhart and Weekley (2006), and Johnson and Oswald (2010), and confirm that with the methods used in the current doctoral thesis, it is possible to develop SJTs that above all reflect the learning outcomes instead of a variety of different constructs. As such, correctly constructed SJTs appear to be a good measurement tool for the EDF in evaluating the domain-specific knowledge of future reserve officers.

In addition, in the conceptual model the covariance between expertise and secondary group cohesion was allowed ($r = 0.65$, $p < 0.001$). Looking at some other significant correlations in the model, it appeared that secondary group cohesion was also positively correlated with extrinsic motivation ($r = 0.45$, $p < 0.001$), and expertise was correlated with extrinsic motivation ($r = 0.31$, $p < 0.001$). Keeping all this in mind, the overall model seems to explain 66% of the variance of intrinsic motivation, which however did not have a statistically significant influence on the results of the SJTDM. Nevertheless, these findings seem to demonstrate that the higher commitment to the EDF and internalization of its organizational values promotes intrinsic motivation of the soldiers, which in turn supports the overall process of gaining the expertise, which in turn is essential for good results in the SJTDM.

6. CONCLUSIONS

6.1. Summary of main conclusions

The main conclusions in **Article I answering research question 1.** A problem in the military context could be taken as an unknown entity between the current state and goal state in some practical situation, which, at the same time, is part of a "bigger picture". There are three characteristics that are specific to the definition of a problem in the military sphere, but not so common in any other fields: problem scope (divides into two subcategories: size of group and degree of danger), relational time, and level of significance (Article **I**).

The main conclusions in **Article II answering research question 2.** Instruments measuring decision-making skills in critical, battle-like situations can be generally categorised into three types: (i) live performance (i.e. actual performance-based measurement); (ii) simulated performance (i.e. simulated performance-based measurement); and (iii) tested performance (i.e. tested performance-based measurement). Of these, tested performance types of instruments seem to be optimal for measuring military tactical-level decision-making skills. It was found that SJT type of tests meet the criteria the best, due to the fact that they were found to be simple, flexible and cost-effective, and yet reliable measures of decision-making skills, making it easy to adapt them to different situations.

The main conclusions in **Article III answering research question 3.** It was found that the methods used for testing development enabled to compile the tests, which had already good quality indicators and thus enabled differentiation between respondents with different levels of military expertise and experience. However, based on the knowledge gained through this doctoral thesis, it is possible to mention that the methods for developing the tests can be further improved for future studies. It also appeared in this doctoral thesis that the quality of such tests can be improved even further based on the empirical data. This improvement can be done very easily by changing the pre-defined scoring schema as a result of one parameter IRT analysis. This in fact can be done every time new data becomes available, and if needed, old values (based on old data) can also be recalculated. Together with the fact that SJTs are relatively easy to compile, it can be thus concluded that instruments developed in this doctoral thesis can be used for assessing military decision-making skills on infantry platoon level.

The main conclusions in **Article IV answering research question 4**. It appeared that more experienced test takers (cadets) achieved better results in the SJTDM than less experienced (conscripts), and thus it was further confirmed that experience is an important characteristic in solving tests, and the developed test is suitable for recognizing these differences. Nevertheless, there were no differences in test results based on the ways the situation was presented (paper or video). However, those test takers who had the situation presented on paper were able to identify slightly more problems in the situations than those who had the situation presentation in video format. On the contrary, those test takers who watched the situation presentation as a video were able to solve the tests slightly faster than those who had the situation presentation in paper format.

The main conclusions in **Article V answering research question 5**. It appeared that only a handful of factors, other than the level of military training and experience (expertise), had a small predictive impact on the results of the SJTDM in our conceptual model (Article **V**). The SJTDM appears to be a good tool for the EDF in evaluating the domain-specific knowledge of future reserve officers. Finally, it appears that by using the methodology developed in this doctoral thesis, it is possible to develop SJTs which above all reflect the learning outcomes instead of a variety of different constructs.

6.2. Limitations of the doctoral thesis

Phase 1. While looking for suitable instruments, many interesting tests relevant to this doctoral thesis were found not in highly regarded scientific journals, but instead in practically oriented army reports. In those cases, we were able to discover those instruments only in the case of some part of the study that was in details described in military reports, was also published in some journal, or we just chased it down through the references in a second search round of literature review. Therefore, it could be that many interesting studies were overlooked, which is a main limitation of phase 1.

Phase 2. One limitation of phase 2 is that the current test did not include cross validation with another sample or a measure of predictive validity, and it also lacks criterion-related validity evidence. Another limitation is the rather small sample size. Both limitations are interconnected, because for validity tests, usually another sample or different set of similar data for the same sample is needed. Obtaining similar type of data for the current sample (for checking predictive validity) is very complicated in the case of the EDF due to different reasons. Firstly, in the best case, this type of data can be gathered during the exercises. However, there are no procedures or valid measure in the EDF to obtain this type of data, not to mention that cadets and conscripts do not attend similar exercises at the same time. It is also very resource intensive to gather such data in a reliable way during the exercises, and thus it requires a lot of planning and preparation, if ever attempted. This exceeded the volume of the current doctoral thesis and thus it remains to be addressed in the future. The second best way to get data for predictive validity is to ask the opinion from the commanders about the decision-making skills of each conscript and cadet tested. This is also difficult, because due to the complicated wartime (reserve army) positioning system, it is extremely difficult to even find out who are the commanders (requires permission to access secret documents via special procedures) of the future platoon leaders, etc. And it is likely that those commanders, even if cleared, would have never seen their subordinate commanders in action. It is also complicated due to the fact that most of the cadets do not have war time positions as platoon commanders before they finish military academy, and after that their war time positions will not be at the same level as conscripts. Thus it is almost impossible to get the data from commanders, not to mention that it would be very subjective anyway, if gathering was ever attempted. Finally, it is possible to use some sorts of grades for predictive validity; however, cadets and conscripts do not attend the same courses, and in the case of cadets, they attend the same courses, but in different years. Thus, in the case of the EDF, it is not possible to get the same data for the current sample, not to mention that the grades given by different teachers are potentially subjective, and thus probably not reliable.

The possibility to get another and/or bigger comparable sample for conducting other validity checks is once again very difficult, because the EDF is very small and it takes a lot of time to introduce a new appropriate sample. The sample available for this doctoral thesis represents all of the cadets at the ENDCOL in 2017, and all the conscripts attending the reserve platoon leader course in 2018. ENDCOL enrols a maximum 50 new land force cadets annually, and 80 new conscripts attend the reserve platoon commander course each year.

Phase 3. One limitation of phase 3 concerns test taking procedures, which were carried out during this doctoral thesis. For example, the time it took to complete the tests was almost two hours (including introduction and all preparations) and that could have affected the test takers' motivation. Another limitation could be small differences in test taking, due to the fact that tests were carried out in small groups, and small differences were inevitable. This effect could have been further increased by the fact that one of the people conducting the tests was changed in the middle of tests, which means that cadets and conscripts might have had slightly different conditions during test taking. For most of the conscripts, tests were conducted quite late in the evening, which means that they could already have been tired from the routine activities of the day, which could have also affected their results.

Phase 4. Though a good conceptual model was found, based on theoretical foundation and data available, only a handful of predicting variables were used. It is thus a limitation of phase 4, because there might be other theories, and thus other factors, which might predict participants' test-taking results better. Other theories and instruments (for example personality traits, abilities, etc.) remain to be tested in the future.

6.3. Implications

6.3.1. Theoretical implications

Phase 1. On the basis of the results of the phase 1, it appears that there is a need for a practical classification of instruments to measure military tactical decision-making skills. In this doctoral thesis, a classification with three types of measurement instruments was offered: (1) actual performance-based measurement, (2) simulated performance-based measurement, and (3) tested performance-based measurement.

Phase 2. The methods used in phase 2 of the doctoral thesis may serve as practical example towards how such tests can be developed and improved even further based on the empirical data. Together with the fact that SJTs are relatively easy to compile, it means that based on the findings of phase 2, it is also possible to check or improve the instrument every time new data becomes available, and thereafter also apply the improvements to the older data if needed. It is also important to point out that during the process of test construction, much attention should be given to framing answer options, which pays off later when administering the results. As a practical recommendation, the EDF should consider working out guidelines for doing that.

Phase 3. The SJTDM could be used as an instrument in other scientific studies with an aim to find out what changes and interventions would be needed to make training of reserve officers more effective.

Phase 4. In the conceptual model, only a handful of factors (other than expertise) had a small influence on the SJTDM results, indicating that it is possible to measure only specific learning outcomes. The findings thus confirm that if the aim is to develop the tool for objective evaluation, then by using the methods developed in the current doctoral thesis, it is possible to compile SJTs that above all reflect the learning outcomes, instead of a variety of different constructs.

6.3.2. Practical implications

Phase 1. The results of the literature review may be useful for researchers and practitioners from other countries, if they face the need to find or adopt suitable instruments for different purposes in improving or testing decision-making skills in a low level military leadership context. Literature reviews should be considered very useful in the case of uncertainty about what has been done in the past that is scientifically sound. For future considerations; however, it is recommended that in the case of literature reviews in such a practical field as the military profession, more attention should be devoted to locally published reports, in addition to database-based literature searches, which might give limited results.

Phase 2. The process of developing the tests was described in detail, which enables duplication by anybody if needed. However, it is recommended for other researchers to carry out validity checks (cross-validation or/and to test predictive validity) of the tests with different samples and in different contexts – for example in other countries if possible. Also, issues with criterion related validity (other than discrimination) remain for further investigation.

Phase 3. The instrument developed in the current doctoral thesis appears to have a number of applications in the EDF, and potentially in other military institutions in different countries; for purposes such as conducting military training, assessment of training outcomes, and recognizing training gaps and needs. The instrument may also have a plethora of possible applications in different spheres of military education, including military studies. For example, they can be used by cadets and students in ENDCOL for the purpose of their bachelor's or master's degree studies. In the case of ENDCOL, it is recommended to consider to carry out further studies in order to find out how well the instrument predicts cadets' academic abilities and overall success of their studies. In addition, this instrument can be used for supporting personnel selection processes (as an additional possibility to gather job related information about the candidates) before or after different stages of training, for example, in the EDF reserve platoon leader course.

Phase 4. The findings enable the EDF and other armies to develop instruments that enable evaluation of military personnel learning outcomes cost effectively yet still objectively (without much interference). Based on the results, it is possible to consider redesigning training in a way that makes it more effective. The findings also enable further practical scientific studies; for example, it should be possible to measure how much domain specific knowledge will be lost during the time after conscription and the first reservist exercise.

7. REFERENCES

Ahronson, A. & Cameron, J. (2007). The Nature and Consequences of Group Cohesion in a Military Sample. *Military Psychology, 19*(1), 9–25. doi:https://doi.org/10.1080/08995600701323277

Allik, S. & Talves, K. (2016). Inimressursi ja kompleksuuringu väljatöötamine kaitsevaldkonnas. *Inimressurss ja riigikaitse: Tervis*, Tartu.

Anderson, J. R. (2015). Chapter 8: Problem Solving. In J. R. Anderson, *Cognitive Psychology* (pp. 181–219). New York: Worth Publishers.

Arlin, P. K. (1989). The problem of the problem. In J. Sinnott & S. J (Ed.), *Everyday problem solving: Theory and applications* (pp. 229–237). New York: Praeger.

Arlin, P. K. (1989). The Problem of the Problem. In S. J. Sinnott, *Everyday Problem Solving: Theory and Applications* (pp. 229–237). New York: Praeger.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. doi:10.1037/0033-295X.84.2.191

Bandura, A. (1982). Self-efficancy mechanism in human agency. *American Psychologist, 37*(2), 122–147. doi:10.1037/0003-066X.37.2.122

Bartone, P. T. & Kirkland, F. R. (1991). Optimal Leadership in Small Army Units. In R. Gal & A. D. Mangelsdorff, *Handbook of Military Psychology* (pp. 393–409). West Sussex: Wiley.

Bartram, D. (2007). Increasing Validity with Forced-Choice Criterion Measurement Formats. *International Journal of Selection and Assessment, 15*(3), 263–272. doi:10.1111/j.1468-2389.2007.00386.x

Beal, D. J., Cohen, R. R., Burke, M. & McLendon, C. L. (2003). Cohesion and Performance in Groups: A Meta-Analytic Clarification of Construct Relations. *Journal of Applied Psychology*, 989–1004. doi: https://doi.org/10.1037/0021-9010.88.6.989

Bennet, A., Bennet, D. H. & Avedisian, J. (2015). *The course of Knowledge: A 21st Century Theory.* Frost, West Virginia: MQIPress.

Bennett, A. & Bennett, D. (2014). Knowledge, Theory and Practice in Knowledge Management: Between Associative Patterning and Context-Rich Action. *Journal of Entrepreneurship, Management and Innovation, 10*(4), 1–41.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B. & Juraska, S. E. (2006). Scoring Situational Judgment Tests: Once You Get the Data, Your Troubles Begin. *International Journal of Selection and Assessment, 14*(3), 223–235. doi:10.1111/j.1468-2389.2006.00345.x

Berisha, G., Pula, J. S. & Krasnigi, B. (2018). Convergent validity of two decision making style measures. *Journal of Dynamic Decision Making, 4*(1), 1–8. doi:10.11588/jddm.2018.1.43102

Brown, A. (2012). Fitting a Thurstonian IRT Model to Forced-Choice Data Using Mplus. *Behavior Research Method*, 1135–1147.

Buch, R., Säfvenbom, R. & Boe, O. (2016). The relationships between academic self-efficacy, intrinsic motivation, and perceived competence. *Journal of Military Studies, 6*(1), 19–35. doi:https://doi.org/10.1515/jms-2016-0195

Byrne, B. M. (2016). *Structural Equation Modelling with AMOS: Basic Concepts, Applications, and Programming (3rd ed.).* New York: Routledge.

Campbell, S. J. (2006). Chapter I. What Has Befallen Me? The Psychological Aftermath of Combat. In T. W. Britt, C. A. Castro & A. B. Adler, *Military Life. The Psychology of Serving in Peace and Combat Military Performance, Volume I* (pp. 3–12). Westport, Connecticut, London: Praeger Security International.

Champion, M., Ployhart, R. & MacKenzie, W. I. (2014). The State of Research on Situational Judgment Tests: A Content Analysis and Directions for Future Research. *Human Performance, 27*(4), 283–310. doi:10.1080/08959285.2014.929693

Chan, D. & Schmitt, N. (1997). Video-Based Versus Paper and Pencil Method of Assessment in Situational Judgment Tests: Subgroup Differences in Test Performance and Face Validity Perceptions. *Journal of Apllied Psychology, 82*(1), 143–159. doi:https://doi.org/10.1037/0021-9010.82.1.143

Chan, D. & Schmitt, N. (2005). Situational judgment tests. In A. Evers, O. Smit-Voskuijl & N. Anderson, *Handbook of personnel selection* (pp. 219–242). Oxford, UK: Blackwell.

Chancey, E. & Bliss, J. (2012). Unreliable Information in Infantry Situation Awareness: Improvement Through Game-Based Training. *Simulation & Gaming, 43*(5), 581–599. doi:10.1177/1046878112439452

Cheung, M. W. & Chan, W. (2002). Reducing Uniform Response Bias with Ipsative Measurement in Multiple-Group Confirmatory Factor Analysis. *Structural Equation Modeling*, 55–77.

Christian, M., Edwards, B. & Bradley, J. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83–117. doi:10.1111/j.1744-6570.2009.01163.x

Cohen, L. L. (2007). *Research Methods in Education* (6 ed.). New York: Taylor & Francis.

Cohen, M. S. & Thompson, B. B. (1999). *Training Teams to Take Initiative: Critical Thinking in Novel Situations.* Arlington: Cognitive Technologies, Inc.

Cohen-Hatton, S. & Honey, R. (2015). Goal-oriented Training Affects Decision-Making Processes in Virtual and Simulated Fire and Rescue Environments. *Journal of Experimental Psychology: Applied*, 395–406.

Connelly, M., Gilbert, J., Zaccaro, S., Threlfall, K., Marks, M. & Mumford, M. (2000). Exploring the Relationship of Leader Skills and Knowledge to Leader Performance. *Leadership Quarterly*, 65–86.

Connolly, T., Arkes, H. & Hammond, K. (2000). *Cambridge series on judgment and decision making. Judgment and decision making: An interdisciplinary reader (2nd ed.).* New York, NY, US: Cambridge University Press.

Coutinho, S. (2008). Self-efficacy, metacognition, and performance. *North American Journal of Psychology, 10*(1), 165–172.

Davis, G. A. (1973). *Psychology of problem solving: Theory and practice.* New York: Basic Books.

Davis, G. A. (1973). *Psychology of Problem Solving: Theory and practice.* New York: Basic Books.

Deci, E. L. & Ryan, R. M. (2000). The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry, 11*(4), 227–268. doi: https://doi.org/10.1207/S15327965PLI1104_01

Dees, R., Nestler, S. & Kewley, R. (2013). WholeSoldier Performance Appraisal to Support Mentoring and Personnel Decisions. *Decision Analysis*, 82–97.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *The Journal of Human Factors and Ergonomics Society*, 32–64.

Endsley, M. & Jones, W. (1997). *Situation awareness information dominance and information warfare (AL/CF-TR-1997-0156).* Wright-Patterson Air Force Base, OH: Crew Systems Directorate, Human Engineering Division.

Eysenck, M. W. (1984). *A handbook of cognitive psychology.* London: Hillsdale (N.J): Erlbaum.

Foldes, H., Ferro, G., Vasilopulous, N., Cullen, M., Wisecarver, M. & Beal, S. A. (2010). *Assessing Judgment Proficiency in Army Personnel.* Arlington: Personnel Decisions Research Institutes.

Frederico, P. (1997). An Empirical Examination of Metacognitive Models of Situation Assessment. *Human Factors, 39*(1), 149–157. doi:10.1518/001872097778940650

Fuglseth, A. M. & Grønhaug, K. (1995). Task Characteristics and Expertise. In G. Kaufmann, T. Helstrup & K. H. Teigen, *Problem solving and cognitive processes* (pp. 513–529). Bergen: Fagbokforlaget.

Gagné, M. & Deci, E. L. (2005). Self-Determination Theory and Work Motivation. *Journal of Organizational Behavior, 26*(4), 331–362. doi:https://doi.org/10.1002/job.322

Gagné, M., Forest, J., Gilbert, M.-H., Aubé, C., Morin, E. & Malorni, A. (2010). The Motivation at Work Scale: Validation Evidence in Two Languages. *Educational and Psychological Measurement*, 628–646.

Glaser, R. (1985). *The Nature of Expertise. Occasional Paper No. 107.* The Ohio State University, The National Center for Research in Vocational Education. Ohio: Columbus.

Gorman, J. C., Cooke, N. J. & Winner, J. L. (2006). Measuring Team Situation Awareness in Decentralized Command and Control Environments. *Ergonomics*, 131213–25.

Grier, R. (2012). Military Cognitive Readiness at the Operational and Strategi Levels: A Theoretical Model for Measurement Development. *Journal of Cognitive Engineering and Decision Making*, 358–392.

Griffith, J. (2006). What Do the Soldiers Say?: Needed Ingredients for Determining Unit Readiness. *Armed Forces and Society, 32*(3), 367–388. doi:https://doi.org/10.1177%2F0095327X05280555

Gully, S. M., Devine, D. J. & Whitney, D. J. (1995). A meta-analysis of cohesion and performance: Effects of levels of analysis and task interdependence. Small Group Research, 26(4), 497–520. doi:https://doi.org/10.1177/1046496495264003

Gully, S., Dennis, J. & Whitney, D. (2012). A Meta-Analysis of Cohesion and Performance: Effects of Level of Analysis and Task Interdependence. *Small Group Research, 43*(6), 702–725. doi:https://doi.org/10.1177/1046496412468069

Hale, K., Stanney, K. & Malone, L. (2009). Enhancing Virtual Environment Spatial Awareness Training and Transfer through Tactile and Vestibular Cues. *Ergonomics, 52*(2), 187–203. doi:10.1080/00140130802376000

Hastie, R. (2001). Problems for Judgment and Decision Making. *Annual Review of Psychology, 52*, 653–683. doi:https://doi.org/10.1146/annurev.psych.52.1.653

Hauenstein, N. M., Findlay, R. A. & McDonald, D. P. (2010). Using Situational Judgment Tests to Assess Training Effectiveness: Lessons Learned Evaluating Military Equal Opportunity Advisor Trainees. *Military Psychology*, 262–281.

Hedlund, E. (2019). A Generic Pedagogic model for Academically Based Professional Officer Education. *Armed Forces & Society, 45*(2), 333–348. doi:10.1177%2F0095327X17749488

Hedlund, J., Horvath J, A., Forsythe G, B., Snook, S., Bullis R, C., Dennis, M. & Sternberg R, J. (1998). *Tacit Knowledge in Military Leadership: Evidence of Construct Validity.* Alexandria: United States Army Research Institute for the Behavioral and Social Sciences. Technical Report 1080.

Horvath, J. A., Sternberg, R. J., Forsythe, G. B., Sweeney, P. J., Bullis, R. C., Williams, W. M. & Dennis, M. (1996). *Tacit Knowledge in Military Leadership: Supporting*

Instrument Development. New Haven: United States Army Research Insitute for the Behavioral and Social Sciences.

Ivey, G. W. & Kline, T. J. (2009). Transformational and Active Transactional Leadership in the Canadian Military. *Leadership & Organizational Development Journal, 31*(3), 246–262. doi:https://doi.org/10.1108/01437731011039352

Jacobs, T. O. (1991). Introduction to Section 4. In R. Gal & D. A. Mangelsdorff, *Handbook of Military Psychology.* West Sussex: John Wiley & Sons Ltd.

Johnson, J. W. & Oswald, F. L. (2010). Test administration and the use of test scores. In J. L. Farr & N. T. Tippins, *Handbook of employee selection* (pp. 151–169). New York: Routledge/Taylor&Francis Group.

Jonassen. (2012). Designing for Decision Making. *Educational Technology Research and Development*, 341–359.

Jonassen, D. (2000). Toward a design theory of probklem solving. *Educational Technology Research and Development, 48*(4), 63–85. doi:https://doi.org/10.1007/BF02300500

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*, 63–85.

Kaju, G. (2013). *Aruanne kaitseväekohustuse täitmisest ja kaitseväeteenistuse korraldamisest 2013. aastal.* Tallinn: Kaitseministeerium.

Karton, I., Männiste, T., Tepp, A. & Kornilov, I. (2018). Ajateenijate teenistusmotivatsiooni ja teenistuse jäämise soovi prognoosivad tegurid. *Estonian Journal of Military Studies*, 138–169.

Keebler, J., Jentsch, F. & Schuster, D. (2014). The Effects of Video Game Experience and Active Stereoscopy on Performance in Combat Identification Tasks. *Human Factors The Journal of the Human Factors and Ergonomics Society, 56*(8), 1482–1496. doi:10.1177/0018720814535248

King, A. (2006). The Word of Command: Communication and Cohesion in the Military. *Armed Forces & Society, 32*(4), 493–512. doi:https://doi.org/10.1177/0095327X05283041

Kirke, C. (2009). Group Cohesion, Culture, and Practice. *Armed Forces & Society*, 745–753. doi:https://doi.org/10.1177/0095327X09332144

Kirke, C. (2010). Military Cohesion, Culture, and Social Psychology. *Defense and Security Analysis*, 143–159. doi:https://doi.org/10.1080/14751798.2010.488856

Klein, G. & Klinger, D. (1991). Naturalistic Decision Making. *Human Systems IAC Gateway*, 16–19.

Klein, G. & Klinger, D. (2000). Naturalistic Decision Making. *Gateway, 11*(3), 16–19.

Kusurkar, R. A., Ten Cate, T. J., Vos, C. M., Westers, P. & Croiset, G. (2012). How motivation affects academic performance: a structural equation modelling analysis. *Advances in Health Sciences Education, 18*, 57–69. doi:https://doi.org/10.1007/s10459-012-9354-3

Kvale, S. (2006). Dominance Through Interviews and Dialogues. *Qualitative Inquiry, 12*(3), 480–500.

Käihko, I. (2018). Introduction to the Armed Forces & Society Forum on Broadening the Perspective on Military Cohesion. *Armed Forces & Society, 44*(4), 563–570. doi:https://doi.org/10.1177/0095327X18763154

Käihko, I. & Haldén, P. (2020). Full-Spectrum Social Science for a Broader View on Cohesion. *Armed Forces & Society, 46*(3), 517–522. doi:https://doi.org/10.1177/0095327X19841669

Legree, P. J. & Psotka, J. (2006). *Refining Situational Judgment Test Methods.* Arlington: U.S. Army Research Institute for the Behavioral and Social Sciences .

Lievens, F. & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *91*(5), 1181–1188. doi:https://doi.org/10.1037/0021-9010.91.5.1181

Lincoln, Y. S. & Guba, E. (1985). *Naturalistic Inquiry.* Beverly Hills, CA: Sage.

Lussier, J. & Shadrick, S. (2004). *Adaptive Thinking Training for Tactical Leaders.* Alexandria: U.S. Army Research Institute – Fort Knox Armored Forces Research Unit.

MacCoun, R. J., Kier, E. & Belkin, A. (2006). Does Social Cohesion Determine Motivation in Combat?: An Old Question with an Old Answer. *Armed Forces & Society, 32*(4), 646–654. doi:https://doi.org/10.1177/0095327X05279181

Macgregor, D. A. (1992). *Future Battle: The Merging Levels of War.* Carlisle, PA, US: U.S. Army War College.

Mael, F. A. & Alderks, C. E. (1993). Leadership Team Cohesion and Subordinate Work Unit Morale and Performance. *Military Psychology, 5*(3), 141–158. doi:https://doi.org/10.1207/s15327876mp0503_1

Marsh, H. W., Hau, K.-T. & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle, *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Lawrence: Erlbaum Associates Publishers.

Mayer, R. & Wittrock, M. (1996). Problem-solving transfer. In D. C. Berliner & R. Calfee, *Handbook of educational psychology* (pp. 47–62). New York: Macmillan.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L. & Grubb, W. (2007). Situational Judgment Tests, Response Instructions, and Validity: a Meta-Analysis. *Personnel Psychology, 60*(1), 63–91.

McDaniel, M., Morgeson, F., Finnegan, E., Campion, M. & Braverman, E. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology*, 730–740.

Meacham, J. & Emont, N. (1989). The Interpersonal Basis of Everyday Problem Solving. In J. Sinnott (Ed.), *Everyday problem solving: Theory and applications* (pp. 7–23). New York: Praeger.

Means, B., Salas, E., Crandall, B. & Jacobs, T. (1993). Training decision makers for the real world. In G. Klein, J. Orasanu, R. Calderwood & E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 306–326). Norwood, NJ: Ablex.

Meerits, A. (2012). *Eesti Kaitseväes kasutatavad juhtimisstiilid ja nende seos soorituse ning sidususega*. Tartu.

Mohdin, A. (2018, July 3). *European countries are quietly bringing back the draft.* Retrieved September 4, 2019, from QUARTZ: https://qz.com/1318379/france-joins-sweden-and-lithuania-in-bringing-back-mandatory-national-service/

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. & PRISMA, G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*.

Morrison, J. E. & Fletcher, J. D. (2002). *Cognitive Readiness.* Alexandria, Virginia: Institute for Defense Analyses.

Motowildo, S., Dunnette, J. & Carter, G. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640–647.

Mullen, B. & Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin, 115*(5), 210–227. doi:https://doi.org/10.1037/0033-2909.115.2.210

Nesbitt, P., Kennedy, Q., Alt, J. & Fricker, R. (2015). Iowa Gambling Task Modified for Military Domain. *Military Psychology, 27*(4), 252–260. doi:10.1037/mil0000080

Nitko, A. (2001). *Educational Assessment of Students. Third Edition.* Upper Saddle River, NJ: Merrill.

Norman, G. (2006). Building on experience: The development of clinical reasoning. *New England Journal of Medicine*, 2251–2252.

Nunnally, J. C. (1978). *Psychometric Theory (2nd ed.).* New York: McGraw-Hill.

Oliver, L., Harman, J., Hoover, E., Hayer, S. & Pandhi, N. (1999). A Quantitative Integration of the Military Cohesion Literature. *Military Psychology, 11*(1), 57–83. doi:https://doi.org/10.1207/s15327876mp1101_4

Patterson, R. E., Pierce, B. J., Bell, H. H. & Klein, G. (2010). Implicit Learning, Tacit Knowledge, Expertise Development, and Naturalistic Decision Making. *Journal of Cognitive Engineering and Decision Making, 4*(4), 289–303. doi:10.1177/155534341000400403

Pawiński, M. (2018). Unintended Consequences of Military Cohesion. *International Peacekeeping, 25*(2), 293–313. doi: https://doi.org/10.1080/13533312.2017.1412261

Ployhart, R. E. & Ehrhart, M. G. (2003). Be Careful What You Ask For: Effects of Response Instructions on the Construct Validity and Reliability of Situational Judgment Tests. *International journal of Selection and Assessment, 11*(1), 1–16.

Ployhart, R. E. & Weekley, J. (2006). Situational judgment: Some suggestions for future science and practice. In J. Weekley & R. PLoyhart, *Situational judgement tests* (pp. 261–264). Mahlwah, NJ: Lawrence Erlbaum Associates.

Rasmussen, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance. *IEEE Transations on Systems, Man, and Cybernetics*, 257–266.

Richman-Hirsch, W. L., Olson-Buchanan, J. & Drasgow, F. (2000). Examining the Impact of Administration Medium on Examinee Perceptions and Attitudes. *Journal of Applied Psychology, 85*(6), 880–887. doi:https://doi.org/10.1037/0021-9010.85.6.880

Robertson, S. (2001). *Problem Solving.* Hove: Psychology Press.

Rowe, A. J. & Mason, R. O. (1987). *Managing with style: A guide to understanding, assessing, and improving decision making.* San Francisco, CA, US: Jossey-Bass.

Ryan, R. M., Deci, E. L. & Edward, L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.

Saaty, T. L. (2008). Decision Making with the Analytic Hierarchy Process. *International Journal of Services Sciences*, 83–98.

Sadler-Smith, E. & Shefy, E. (2004). The Intuitive Executive: Understanding and Applying 'Gut Feel' in Decision-Making. *Academy of Management Executive, 18*(4), 76–91. doi:https://doi.org/10.5465/AME.2004.15268692

Saini, S. K. (2008). Role of Intuition in Military Command. *Journal of Defence Studies*, 75–88.

Salas, E., Priest, H. A., Wilson, K. A. & Burke, S. C. (2006). Scenario-Based Training: Improving Military Mission Performance and Adaptability. *Military life: The psychology of serving in peace and combat: Operational stress*, 32–53.

Saldago, J., Viesewaran, C. & Ones, D. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. Anderson, D. Ones, H. Sinagil & C. Visewaran, *Handbook of industrial, work & organizational psychology, Vol. 1.* (pp. 165–199). London: Sage.

Salmon, P., Walker, G., Baber, C., Jenkind, D., McMaster, R. & Young, M. (2008). What really is going on? Review of situation awareness models for individuals. *Theoretical Issues in Ergonomics, 9*(4), 297–323. doi:10.1080/14639220701561775

Salo, M. (2006). Beyond Training Alone: The Role of Cohesion Maximizing Group Performance. *Tiede ja Ase, 64*, 160–174.

Salo, M. & Siebold, G. L. (2007). The Structure of Military Cohesion: Components, Predictors, and Outcomes. In H. Annen & W. Royl, *Military pedagogy in progress. Studies for Military Pedagogy, Military Science & Security Policy* (Vol. 10, pp. 213–224). Frankfurt am Main, Germany: Peter Lang.

Salo, M. & Siebold, G. L. (2008). Variables impacting peer group cohesion in the finnish conscript service. *36*(1), 1–18.

Saner, L., Bolstad, C., Gonzalez, C. & Cuevas, H. (2009). Measuring and Predicting Shared Situation Awareness in Teams. *Journal of Cognitive Engineering and Decision Making*, 280–308.

Saus, E., Johsen, B., Eid, J. & Thayer, J. (2012). Who benefits from simulator training: Personality and heart rate variability in relation to situation awareness during navigation training. *Computers in Human Behavior, 28*(4), 1262–1268. doi:10.1016/j.chb.2012.02.009

Scott, S. G. & Bruce, R. A. (1995). Decision-Making Style: The Development and Assessment of a New Measure. *Educational and Psychological Measurement, 55*(5), 818–831. doi:10.1177/0013164495055005017

Shadrick, S. & Lussier, J. (2004). *Assessment of the Think Like a Commander Training Program.* Fort Knox: U.S. Army Research Institute for the Behavioural and Social Sciences.

Shamir, B., Brainin, E., Zakay, E. & Popper, M. (2000). Perceived Combat Readiness as Collective Efficacy: Individual and Group-Level Analysis. *Military Psychology, 12*(2), 105–119. doi:https://doi.org/10.1207/S15327876MP1202_2

Shils, E. A. & Janowitz, M. (1948). Cohesion and Disintegration in the Wehrmach in World War II. *Publich Opinion Quarterly*, 280–315.

Siebold, G. L. (2006). Military Group Cohesion. In T. W. Britt, C. A. Castro & A. B. Adler, *Military Life: the Psychology of Serving in Pease and Combat. Volume 1: Group Performance* (pp. 185–201). Westport, Connecticut, London: Praegar Security International.

Siebold, G. L. (2007). The Essence of Military Cohesion. *Armed Forces & Society*, 286–295.

Siebold, G. L. (2007). The Essence of Military Cohesion. *Armed Forces & Society, 33*(2), 286–295. doi:https://doi.org/10.1177%2F0095327X06294173

Siebold, G. L. (2011). Key Questions and Challenges to the Model of Military Group Cohesion. *Armed Forces & Society*, 448–468.

Siebold, G. L., Crabb, T., Woodward, R. & King, A. C. (2016). Combat, Cohesion, and Controversy: Disputatio Sine Fine. *Armed Forces & Society, 42*(2), 449–462. doi:https://doi.org/10.1177/0095327X15573213

Sinnott, D. J. (1989). *Everyday problem solving: Theory and applications.* New York: Praeger.

Sohn, Y. & Doane, S. (2004). Memory processes of flight situation awareness: Interactive roles of working memory capacity, long-term working memory, and expertise. *Human Factors, 46*(3), 462–474. doi:10.1518/hfes.46.3.461.50392

Sookermany, A. (2012). Learning in Doing – Skills Acquisition in [Post-] Modernised Military. *Defence Studies, 11:4*, 615–635.

Spiker, A. & Johnston, J. (2013). Assessing Counter-Terrorism field with Multiple Behavioural Measures. *Applied Ergonomics*, 690–686.

Stajkovic & Luthans. (1998). Self-Efficacy and Work-Related Performance: A Meta-Analysis. *Psychological Bulletin*, 240–261.

Zeichmeister, E. B. & Johnson, J. E. (1992). *Critical thinking: A functional approach.* Pacific Grove, CA: Brooks/Cole.

Tan, K. H., Tse, Y. K. & Chung, P. L. (2010). A plug and play pathway approach to operations management games development. *Computers & Education, 55*(1), 109–117. doi:https://doi.org/10.1016/j.compedu.2009.12.008

Thomas, C. & Hirschfeld, R. (2015). Knowing is half the battle: Interdependent effects of knowledge and action on leader emergence. *Leadership & Organization Development Journal*, 512–526.

Thompson, M. M. & McCreary, D. R. (2006). Enhancing mental readiness in military personnel. In A. Adler, C. Castro & T. Britt, *Military life: The psychology of serving in peace and combat: Vol. 2. Operational stress* (pp. 54–79). Westport, CT: Praeger Security.

Thunholm, P. (2004). Decision-making style: habit, style or both? *Personality and Individual Differences, 36*(4), 931–944. doi:10.1016/S0191-8869(03)00162-4

Thunholm, P. (2009). Military Leaders and Followers – Do They Have Different Decision Styles? *Scandinavian Journal of Psychology*, 317–324.

US.Army. (2010). *The operations process.* Washington: Headquarters, Department of the Army.

Weekley, J. A., Ployhart, R. E. & Holtz, B. C. (2006). On the Development of Situational Judgment Tests: Issues in Item Development, Scaling, and Scoring. In J. A. Weekley & R. E. Ployhart, *Situational Judgment Tests Theory, Measurement, and Application* (pp. 126–143). New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Weick, K., Sutcliffe, K. & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science, 16*(4), 409–421. doi:10.1287/orsc.1050.0133

Whetzel, D. L. & McDaniel, M. A. (2009). Situational Judgment Tests: An Overview of Current Research. *Human Resource Management Review, 19*(3), 188–202.

Vogel-Walcutt, J. J., Carper, T. M., Bowers, C. & Nicholson, D. (2010). Increasing Efficiency in Military Learning: Theoretical Considerations and Practical Applications. *Military Psychology*, 311–339. doi:https://doi.org/10.1080/08995605.2010.492701

Vowell, J. B. (2004). *Between Discipline and Intuition_The Military Decision Making Process in the Army's Future Force.* Fort Leavenworth, Kansas: United Stares Army Command and General Staff College.

8. ANNEXES

Annex 1. Initial and revised (in parentheses after symbol =) score values of SJTDM items (k11-k16, k21-k26, v11-v16, v21-v26) based on the ability mean (AM) of all options of the items used in the IRT analysis and explanations and descriptions of actions taken for revising scoring schema if needed (Article **III**, the table is available at https://sisu.ut.ee/sjt).

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
k11	6	1	7	.58	.17	No changes were needed. Experts' estimate was 1 and it was therefore expected that only a few respondents would select high values. However, quite a few respondents still chose the worst option as the best. This suggests that in addition to information which makes this option a bad choice, there might still be too much information which makes respondents (presumably beginners) consider it a good choice.
	5	2	6	.78	.29	
	4	3	10	1.04	.26	
	3	4	20	1.13	.36	
	2	5	24	1.35	.49	
	1	6	35	1.58	.58	
k12	6	2	3	.55	.36	No changes were needed. Although there were very few respondents who selected options 6 and 5, their ability mean was significantly different from others. Experts' estimate was 2 and, again, it was expected that only a few respondents would select high values. Compared to item k11, it seems, though, that this item was less likely to be considered a good option even though it was closer to them. The substantive reasoning might be that in the case of this item, the test developers included less information, which made respondents consider this option a good or a very good option.
	5	3	3	.71	.37	
	4	4	6	1.01	.45	
	1	5	38	1.23	.43	
	3	5	21	1.35	.55	
	2	6	29	1.47	.59	
k13	6	3	2	.91	.25	Data codes 6 and 5 were merged because of a similar ability mean and low number of answers for both codes; the new score value for these is 3 points (the ability mean difference from the data score awarded with 4 points is about the same as for the data score awarded with 5 points). Experts' estimate was 3 and thus respondents were expected to select 2–4. Choices 5 and 6 had both very few selections and, from a substantive point of view, it might be that there is just enough information in this item for it to be considered close to the experts' estimate. Thus, merging seems reasonable also from a substantive point of view.
	5	4=3	5	.92	.47	
	1	4	18	1.08	.43	
	4	5	12	1.20	.61	
	2	5	30	1.21	.46	
	3	6	33	1.54	.56	

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
k14	1 6 2 5 3 4	3=2 4=3 4=3 5=4 5=4 6	6 8 12 16 15 42	.71 .95 1.18 1.07 1.14 1.57	.30 .38 .51 .39 .39 .55	Data codes 2, 5 and 3 were initially merged because of a similar ability mean; the new score value for these was planned to be 4 points, as the ability mean difference from the most demanding option (code 4) was significantly larger than this in comparison with the code that was one level below (code 6); subsequently, codes 6 and 1 had to be awarded 1 point less. However, in the content analysis it appeared that the characteristics of option 2 were clearly showing that it was worse than options 3 and 5; thus, it would be reasonable to merge it with option 6 and give them both 3 points, although the ability mean of option 2 is higher than for options 3 and 5.
k15	1 2 3 4 6 5	2 3=2 4=3 5=4 5 6	2 4 6 17 30 41	.66 .49 .90 1.02 1.28 1.54	.30 .31 .41 .40 .42 .56	Data code 4 had the same score as data code 6, but the ability mean was lower and the score was therefore lowered; subsequently, the other lower scores were lowered as well, but the score of data code 1 was not lowered because it was merged with data code 2, as in both of these cases, very few respondents selected these options. Experts' estimate was 5; therefore, selections 4–6 were expected. From a substantive point of view, choice 1 and 2 are clearly both very bad, so 2 points for both seems to be justified. In the case of choice 4, lowering the points is also justified because choice 6 is clearly better and thus worth more points.
k16	1 2 3 4 5 6	1=2 2 3=2 4 5 6	1 2 5 13 29 49	.80 .67 .54 1.01 1.23 1.48	.13 .27 .34 .32 .47 .54	Data codes 1, 2 and 3 were merged because of a similar low ability mean and a small number of respondents selecting those options; the new score is 2 because the ability mean differs more from the 5-point code than the 5-point code differs from the 4-point code. Experts' estimate was 6 and it was therefore not expected that very many respondents would choose options 1–3. Even though option 3 seems to be a little bit better than 1 and 2, it might still be reasonable, from a substantive point of view, to merge them all.
k21	6 5 4 3 2 1	1 2 3=4 4 5=4 6	5 12 10 16 17 39	.53 .84 1.16 1.18 1.21 1.60	.19 .28 .40 .28 .32 .61	Data codes 4, 3 and 2 were merged because of a similar ability mean; the new score value for these is 4 points (their average). Experts' estimate was 1 and it was therefore expected that only a few respondents would select high values. However, from a substantive point of view, choices 5 and 6 are clearly not poor options and should therefore both get 2 points. Choices 3 and 4 seem also quite equal and should therefore both get 3 points. Choice 2 is clearly a worse option than choice 3 and should therefore get 4 points.

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
k22	6	2=1	1	.34	.24	The ability mean for data code 2 is significantly more different than the differences between other data codes; therefore, the score of the other ones was lowered; codes 6, 5 and 4 were selected by very few respondents, but their ability means are too different for merging these options. Experts' estimate was 2 and responses 5, 6 and even 4 were therefore not expected to appear often.
	5	3=2	2	.76	.23	
	4	4=3	5	.96	.38	
	1	5=4	38	1.13	.35	
	3	5=4	16	1.22	.53	
	2	6	37	1.54	.61	
k23	6	3=1	1	.65	.14	The ability mean for data code 4 was significantly higher than for data code 2, but the scores were the same; similarly, for data code 5 the ability mean was higher than for data code 1; in order to solve this issue, the score of data codes 2 and 1 were lowered by 1 point and, subsequently, the lower scores were lowered as well. Experts' estimate was 3 and responses between 2 and 4 were therefore expected. However, it seems that more respondents preferred choice 2, which is reasonable because choice 3 includes information which clearly makes it a bad choice. As choice 5 is clearly distinguishable as not a good option, from a substantive point of view, it should also get less points; thus, 2 points seems more reasonable.
	1	4=2	19	.95	.39	
	5	4=3	3	1.07	.46	
	2	5=4	31	1.18	.46	
	4	5	15	1.33	.47	
	3	6	31	1.57	.60	
k24	1	3	2	.85	.35	No changes were needed.
	6	4	12	1.03	.35	
	2	4	9	1.03	.30	
	3	5	30	1.14	.47	
	5	5	18	1.24	.54	
	4	6	29	1.63	.57	
k25	1	2=N/A	1	1.24	.17	Data code 1 was selected by very few respondents, but these appeared to be the ones with a comparatively good total test score; therefore, this option indicated an anomaly and was removed from the analysis; the ability mean of data code 4 was significantly lower than that of data code 6, but the scores were the same; therefore, the score for data code 4 was lowered and, subsequently, other lower scores were lowered as well. Experts' estimate was 5; thus, selections 4–6 were expected to appear more often. From a substantive point of view, there is no good explanation as to why the overall better respondents have chosen this option as the worst. Thus, re-scoring seems reasonable.
	2	3=2	4	.46	.23	
	3	4=3	5	.80	.37	
	4	5=4	33	1.08	.47	
	6	5	10	1.30	.36	
	5	6	47	1.51	.53	

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
k26	1 2 3 4 5 6	1=2 2 3=2 4 5 6	0 2 1 7 18 71	.51 .83 .62 .98 1.06 1.39	.00 .42 .43 .42 .43 .54	Data codes 1, 2 and 3 were merged because they had a very small number of respondents and a comparatively similar ability mean value; the new score is 2, as the ability means are more different from the data code with score 4 than the difference in the case of data code with score 5. Experts' estimate was 6 and it was therefore not expected that very many respondents would choose options 1–3. Even though option 1 is clearly the worst one and distinguishable from 2 and 3, it might still be reasonable, from a substantive point of view, to merge them all.
v11	6 5 4 3 2 1	1=2 2 3 4 5 6	2 7 10 14 28 29	.84 .92 1.01 1.18 1.24 1.56	.40 .39 .43 .44 .52 .56	Data codes 6 and 5 were merged because of a similar ability mean and low number of answers for code 6; the new score value for these is 2 points (the ability mean difference from the data score awarded with 3 points is about the same as for the data score awarded with 4 points). Experts' estimate was 1 and it was therefore expected that only a few respondents would select values 5 and 6. Thus, from a substantive point of view, re-scoring seems reasonable.
v12	6 5 4 1 3 2	2 3=2 4=3 5=4 5 6	1 1 4 61 9 24	.65 .38 .92 1.19 1.35 1.59	.64 .19 .23 .49 .46 .57	The score of code 1 was lowered because it was the same as that of code 1, but its ability mean was lower; subsequently, the score of codes 4 and 5 was lowered as well and code 5 was merged with code 6, as there were very few respondents for both of these options. Experts' estimate was 2 and responses 5 and 6 and even 4 were therefore not expected to appear very often. Other than that, from a substantive point of view, re-coding seems reasonable.
v13	6 1 5 2 4 3	3 4 4 5=4 5 6	22 5 15 15 20 22	1.04 1.14 1.24 1.19 1.34 1.55	.40 .36 .54 .52 .41 .69	Data codes 1, 5 and 2 were merged because of a similar ability mean and low number of answers for code 1; the new score value for these is 4 points (a value between those options that require less ability to answer correctly (code 6) and those that require more ability to answer correctly (code 4). Experts' estimate was 3 and responses between 2 and 4 were therefore expected. It is strange, however, that so many respondents considered this option as the best, although it seems that this option includes pieces of information that are not so correct but might seem reasonable to beginners. Other than that, re-scoring seems reasonable from a substantive point of view.

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
v14	1	3	2	.91	.36	The ability means for data codes 3 and 4 were almost the same; therefore, both of these were awarded 6 points (not 5 points, as the ability mean is higher than for code 5, which was awarded 5 points). Experts' estimate was 4; thus, respondents' selections 3–5 were expected and considered acceptable from a substantive point of view. However, choices 3 and 4 seem quite similar and therefore, from a substantive point of view, re-scoring seems reasonable even though this is the first case where the same amount of points was awarded to an option which is not the option suggested by experts.
	2	4	10	1.05	.37	
	6	4	17	1.19	.42	
	5	5	21	1.27	.57	
	3	5=6	21	1.35	.53	
	4	6	29	1.37	.62	
v15	1	2	1	.03	.52	Data codes 3 and 6 were merged because of similar ability means; the new score value for these is 4 points (a value between those options that require less ability to answer correctly (code 2) and those that require more ability to answer correctly (code 4); there were very few answers for option 1, but it cannot be merged with other codes because of a significant difference in the ability mean. Experts' estimate was 5 and selections 4–6 were therefore expected to appear more often. However, option 3 seems to have almost the same amount of selections, which is strange from a substantive point of view. Other than that, re-scoring seems reasonable.
	2	3	9	1.07	.47	
	3	4	23	1.13	.36	
	6	5=4	21	1.16	.51	
	4	5	21	1.39	.62	
	5	6	24	1.50	.57	
v16	1	1	2	.55	.53	No changes were made. There was only a small number of respondents who selected options 1 and 2, but the ability means for these were significantly different; therefore, it was not reasonable to re-score these options. Experts' estimate was 6 and it was therefore not expected that very many respondents would choose options 1–3. Thus, from a substantive point of view, this item does not need re-scoring either.
	2	2	3	.86	.24	
	3	3	12	.94	.43	
	4	4	16	1.10	.42	
	5	5	31	1.24	.45	
	6	6	37	1.55	.57	
v21	6	1	0	.01	.00	Data codes 4, 3 and 2 were merged because of similar values of ability means and low number of respondents for codes 4 and 3; the new score is 5 because the ability mean is not very different from code 1 with a score of 6. Experts' estimate was 1 and selections 1–3 were therefore expected. Interestingly, in this case, number 1 was chosen overwhelmingly and, from a substantive point of view, the reason seems to be one piece of information inside this item (Motorola case). However, cases 3 and 4 are still clearly worse compared to option 2; thus, from a substantive point of view, awarding 4 points to those options and 5 points to option 2 is justified.
	4	3=4	0	1.17	.00	
	3	4	2	.93	.35	
	2	5	16	1.06	.42	
	1	6	81	1.33	.55	

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
v22	5 4 3 1 2	3 4=5 5 5 6	1 3 15 13 67	.61 1.01 1.06 1.11 1.38	.48 .53 .39 .45 .56	Data code 4 was scored 1 point higher, as it had an ability mean value comparatively similar to data codes 3 and 1 and it was more different from data code 5, which was scored 3 points, than from data code 2, which was scored 6 points. Experts' estimate was 2 and selections 1–3 were therefore expected. In this case, option 2 was chosen most often and this, again, indicates that there is some information inside it which clearly makes it easy to distinguish as second worse option. In this case, option 6 was never selected, which confirms that overall, situation D2 is easier for respondents due to easily distinguishable items. From a substantive point of view, however, this re-scoring seems justified.
v23	6 1 5 4 2 3	3 4=3 4 5=4 5 6	5 3 18 25 7 42	.87 1.04 1.17 1.19 1.26 1.43	.45 .27 .54 .40 .42 .62	Data codes 5 and 4 were merged because of similar ability means; the new score value for these is 4 points, as the ability mean is lower than that of code 2, where it is 5. Subsequently, the score of data code 1 should be lowered and merged with data code 6; merging is appropriate due to a low number of respondents selecting this option. Experts' estimate was 3 and selections 2–4 were expected. Choice 3 was selected most often; however, choices 4 and 5 were also common. From a substantive point of view, it therefore seems that options in the middle, 3, 4 and even 5, are quite similar in the eyes of respondents. Thus, re-scoring seems reasonable.
v24	1 2 6 5 3 4	3 4=3 4 5 5 6	2 6 14 26 27 25	.81 .94 1.08 1.21 1.29 1.56	.18 .41 .39 .43 .49 .69	Data codes 1 and 2 were merged because of very few answers in the case of option 1 and not too different ability mean scores. The new score value for these is 3 points, as data code 6, which is awarded 4 points, has a slightly higher ability mean, but not so much higher, however, that these codes should get only 2 points (compared to the difference in data codes 6 and 5 or 3). Experts' estimate was 4 and selections between 3 and 5 were therefore expected. Choice 4, however, was not most often chosen, which confirms that items between 3 and 5 seem quite similar to respondents. Thus, from a substantive point of view, re-scoring is justified.

Item	Data code*	Score value**	Freq.* **(%)	AM****	SD of AM	Action taken
v25	2	3	2	.72	.46	No changes were made. There is only a small number of respondents who selected option 2, but the ability mean for this was significantly different from data code 3 and, therefore, it was not reasonable to re-score this option. Experts' estimate was 5, and, from a substantive point of view, it did not need re-scoring even though option 6 was chosen almost as often as option 5.
	3	4	12	1.08	.44	
	6	5	13	1.14	.53	
	4	5	36	1.25	.47	
	5	6	37	1.43	.61	
v26	1	1	1	.29	.28	Data codes 3 and 4 were merged because of very few answers in the case of option 3 and similar ability means. The new score is 4; in addition, the score of data code 2 should increase by 1 because it has an ability mean significantly closer to the codes with a score of 4 rather than 1. Experts' estimate was 6 and it was selected most often.
	2	2=3	1	.81	.43	
	3	3=4	2	1.00	.11	
	4	4	11	.95	.42	
	5	5	17	1.20	.48	
	6	6	68	1.38	.55	

*all the response options selected by at least one participant for any item
**points given for each option if selected (if '=' then re-scored as a result of IRT analysis)
***how frequency the particular option has been chosen
****ability mean – the score that shows how difficult the particular option was to the respondents

9. SUMMARY IN ESTONIAN

Sõjaväeliste juhtide otsustamisoskuse hindamine lahingujuhtimist matkivas olukorras

Külma sõja järgsel perioodil (1990–2008) on mitmed Euroopa riigid otsese sõjaohu puudumiselminetanud esmase kaitsevõime, sealhulgas loobunud kohustuslikust ajateenistusest (Hedlund, 2019). Pärast sündmusi Gruusias (2008) ning Ukrainas (2014) on aga taas hakatud sellele tähelepanu pöörama ning mitmed, eeskätt NATO idapiiri riigid, on taastamas ajateenistust ja seeläbi pööramas järjest enam tähelepanu reservväelaste väljaõppele (Mohdin, 2018). Eesti pole küll ajateenistusest loobunud, kuid on sarnaselt paljudele teistele NATO riikidele pööramas suurt rõhku reservväelaste, eriti aga reservülemate ettevalmistamisele.

Iga aasta alustab Eesti Kaitseväes (EKV) teenistust umbes 1000 eelkutse ajateenijat, kes saavad reservülema ettevalmistuse (Kaju, 2013). EKV on väikesearvuline isegi sõjaaegse mehitatuse korral, mistõttu võib konflikti puhkemise korral loota vaid sellele, et ülemad saavad lahingolukordade lahendamisega paremini hakkama kui vastased. See tähendab, et EKV jaoks on väga oluline ülemate, eeskätt reservülemate väljaõpe, mis muuhulgas peab valmistama nad etteotsustamiseks oma üksuse eesotsas kriitilistes lahinguolukordades. Seetõttu on EKV hakanud tähelepanu pöörama inimressursile ning alustanud pikaperspektiivilist projekti eesmärgiga oma inimressurssi uurida ja arendada (Allik & Talves, 2016). Vaatamata erinevustele armeede suuruses on sarnastele seisukohtadele jõutud ka USA armees (Dees, Nestler & Kewley, 2013).

Reservülemate põhitööks on oma üksuse juhtimine lahingusituatsioonides. Oskuslikud ülemad peavad olema suutelised koostama ja vastavalt olukorrale kohandama plaane, mis tagaks oma eeliste maksimaalse ärakasutamise, minimeerides samal ajal vastase poolt tehtavat kahju. See tähendab, et ülemad peavad stressirohkes olukorras suutma situatsiooni hinnata, töödelda lisanduvat uut infot, suhelda inimestega, silma peal hoida plaanide muutumisel ning olukordade arengul (Lussier & Shadrick, 2004). Erinevalt teistest ametitest pole sõjaväelistel juhtidel võimalik rahuajal lahingujuhtimist praktiseerida muul moel, kui õppeolukordades, mis ei pruugi aga tagada piisavat autentsust. Seetõttu võib isegi elukutselistel kaitseväelastel, rääkimata reservülematest, tekkida raskusi saada lõpuni aru sellest, mida sisaldab endast ülema amet ja mida tähendab tõhus üksuse juhtimine lahingus (Sookermany, 2012).

EKV peab seega tagama võimalikult hea ülemate väljaõppe, mille tulemuslikkuse kohta saab järeldusi teha vaid juhul, kui väljaõppe tulemused on mõõdetavad. Paraku ei ole hetkel EKV-s võimalik teaduslikus tähenduses usaldusväärselt mõõta ülemate lahingujuhtimist, mistõttu on vajadus välja töötada selleks sobiv mõõtevahend, mis tagaks kulutõhusa testimise, kuid annaks teaduslikult usaldusväärseid tulemusi. See tähendab, et testi peaks olema võimalik teha klassiruumis, selle läbimine koos ettevalmistuste ja kokkuvõtetega ei tohiks võtta väga palju aega, samuti ei tohiks sellele kuluda palju inim- ja muid ressursse. Samuti ei tohi testi koostamine olla väga keeruline ega ressursimahukas. Teadaolevalt ei ole hetkel olemas testi, mis mõõdaks taktikalise tasandi ülemate otsustamisoskust lahingujuhtimise kontekstis.

Doktoritöö eesmärk on töötada välja instrument, mis võimaldab mõõta otsustamist lahingujuhtimise olukordades ning testida selle instrumendi sobivust ajateenijate ja kadettide näitel. Seejärel selgitada välja, mis prognoosib testi tulemuslikumat lahendamist ajateenijatest (algajad) ja kadettidest (edasijõudnud) koosneva valimi näitel.

Eesmärgi saavutamiseks sõnastati viis uurimisküsimust:

(i) Millised on probleemilahendamise erisused sõjanduses, mil määral need erinevad probleemilahendamise tavapärasest käsitusest?
(ii) Milliseid instrumente on kasutatud lahingus (kriitilises olukorras) otsustamise tulemuslikkuse teaduslikuks mõõtmiseks ja milline neist instrumentidest sobiks kõige paremini mõõtma teaduslikult usaldusväärselt sõjaväeliste juhtide otsustamiseoskust?
(iii) Kuidas kohandada instrumenti EKV tarbeks ja millised on selle instrumendi kvaliteedinäitajad?
(iv) Kui hästi instrument eristab edasijõudnuid algajatest rühma juhtimistasandi lahingujuhtimist matkivas olukorras EKV kadettide (kogenumad) ja ajateenijate (vähemkogenumad) näitel?
(v) Mille alusel on rühma juhtimistasandi sõjaväeliste juhtide otsustusvõime prognoositav EKV kadettide (edasijõudnud) ja ajateenijate (algajad) näitel?

Eesmärkide saavutamiseks jagati doktoritöö neljaks etapiks: (1) ettevalmistused instrumendi koostamiseks, (2) instrumendi koostamine, andmete kogumine ning empiirikale tuginev instrumendi kvaliteedinäitajate väljaselgitamine ning instrumendi parandamine, (3) guppidevaheliste erinevuste väljatoomine tuginedes testi tulemustele, (4) testi tulemuslikumat sooritamist prognoosivate faktorite tuvastamine.

Esimene etapp jagunes kaheks ning selles otsiti vastuseidesimesele ja teisele uurimisküsimusele. Etapi tulemused on kajastatud artiklites **I** ja **II**. Etapi käigus töötati kirjandusega ja anti ülevaade nii probleemi lahendamise erisustest sõjanduses kui ka instrumentidest, millega on varasemalt mõõdetud otsustamist sõjanduses lahingutegevuse juhtimisel. Artikkel **I** keskendus probleemilahendamise erisustele sõjanduses, andmete kogumiseks kasutati poolstruktureeritud intervjuusid ekspertidega. Valimiks olid kuus kaitseväelast (neli vanemohvitseri, üks nooremohvitser ja üks allohvitser). Intervjuu teemad keskendusid probleemi mõiste ja probleemi lahendamise erisustele sõjanduses. Tulemuste analüüsil kasutati nii deduktiivset kui ka induktiivset lähenemist. Andmete korrastamiseks ja analüüsimiseks kasutati programmi NVIVO-10. Artikkel IIkajastab PRISMA juhistest lähtunud süstemaatilise kirjanduse analüüsi tulemusi, mis keskendus sõjandusvaldkonnas lahingujuhtimises kasutatud mõõdikute tuvastamisele lähtudes seatud piirangutest. Kirjanduse analüüsi jaoks kasutati EBSCO andmebaase, lõpliku kirjanduse nimekirja koostamise protseduur koosnes neljast faasist: (i) sõnastati märksõnad ja valiti välja andmebaasid; (ii) valitud märksõnadega otsitud ja leitud allikaid hinnati kahe sõltumatu uurija poolt sobivuse alusel pealkirja tasandil; (iii) eelmises faasis valitud allikaid hinnati samal viisil abstrakti tasandil ja viimaseks (iv) täisteksti tasandil. Lõpuks jäi allikate nimekirja 23 allikat, mille alusel teostati analüüs.

Teine etapp jagunes kolmeks ning selles otsiti vastuseid kolmandale uurimisküsimusele. Esmalt koostati eelmise etapi käigus kogutud informatsioonile tuginedes instrument, mis võimaldaks mõõta madala juhtimistasandi sõjaväeliste ülemate otsustamisoskust lahingujuhtimise olukordades. Instrumendi koostamisel lähtuti Jonasseni (2000) kirjeldatud strateegilise otsustamise probleemi struktuurist, kuna sellist tüüpi probleemid on sõjanduses lahingujuhtimisel kõige tavalisemad (artikkel **I**), sisaldades reaalajas aset leidvaid lahendamist vajavaid kompleksseid sündmusi, mis võivad kiiresti muutuda. Koostatud instrument koosneb seega kaheosalistest probleemolukordadest (esmane- ja jätkuolukord), kus esmane olukord on lihtsam ning jätkuolukord keerukam. Iga testi stsenaariumi esitamiseks oli ette valmistatud kaks võimalust, neid sai esitada nii videona kui paberkujul, sõltumata esitamise viisist sisaldasid mõlemad versioonid täpselt sama informatsiooni. Testid sisaldasid kuute vastusevarianti, mis valideeriti koostamise käigus ekspertmeetodil. Iga vastusevarianti tuli vastavalt olukorra lahendamiseks sobilikkusele hinnata skaalal 1–6, kus 1 – väga halb ja 6 – väga hea, kusjuures igat numbrit sai kasutada vaid korra. Instrumendi koostamise potseduuri on detailselt kirjeldatud artiklites **III** ja **IV**. Paralleelselt instrumendi koostamisega komplekteeriti valim, mille hulka kuulusid 134 KVA kadetti ning 80reservrühmaülemate baaskursuse ajateenijat. Seejärel viidi läbi testimine (andmete kogumine), mida on detailselt kirjeldatud artiklites **III** ja **IV**. Pärast andmete kogumist selgitati välja loodud instrumendi kvaliteedinäitajad, kasutades selleks andmetöötlusprogrammi SPSS Statistics 25. Seejärel viidi läbi üheparameetriline IRT (Item Response Theory) analüüs eesmärgiga leida küsimuste kvaliteedinäitajad ja keerukus ning vajaduselempiirikale tuginedes küsimusi ja nendes sisalduvaid valikvastuseid ka parandada. IRTanalüüs viidi läbi kasutades programmi WinSteps 4.0.1. Analüüsi tulemuste alusel muudeti mõnede vastusevariantide eest antavaid punkte, mis tõstis OPT usaldusväärsust järgnevateks analüüsideks veelgi. Andmestiku kvaliteediga seonduvaid protseduure on detailselt kirjeldatud artiklis **III**.

Kolmas etapp viidi läbi vastamaksneljandale uurimisküsimusele ning see keskendus erinevate gruppide vaheliste erinevuste tuvastamisele, kasutades selleks andmetöötlusprogrammi SPSS Statistics 25. Detailsed tulemused on kajastatud artiklis **IV**, mis oli koostatud eestikeelsena spetsiaalselt EKV tarbeks jamis sisaldab kokkuvõtteid gruppide vahelistest erinevustest (kadetid vs ajateenijad), lähtudes mitte ainult testi tulemustest, vaid ka võimalikest tulemusi mõjutanud välistest faktoritest: testilahendamise ajast, probleemide arvust, samuti stsenaariumi esitamise erinevatest viisidest (video vs paber). Lisaks keskenduti artiklis **IV** ka ettepanekutele ajateenijate õppe tõhustamisel, samuti koostatud instrumendi edasisele kasutamisele EKV-s.

Neljandas etapis otsiti vastust viiendale uurimisküsimusele ning lisaks testide tulemustele baseerus see ka lahendajate poolt täidetud lisaküsimustiku andmetele.

Lisaküsimustik koosnes EKV inimvara uuringus (Allik & Talves, 2016) kasutatud küsimustiku valitud küsimustest, mida on detailselt kirjeldatud artiklis **V**. Lisaküsimustikuga saadud andmeid kasutati testi lahendamise tulemuslikkust prognoosivates kinnitava faktoranalüüsi (CFA) ja struktuurivõrrandite (SEM) mudelites, kasutades selleks programme IBM SPSS 25 Amos Graphics. Protseduurid ja tulemused on detailselt välja toodud artiklis **V**.

Doktoritöö esimese etapi peamiste tulemustena on artiklis **I** välja toodud probleemi ja probleemilahendamise erisused sõjanduses, samuti pakuti välja kolm lisakategooriat sõjandusalase probleemi kirjeldamiseks: probleemi mõjuulatus, suhteline aeg ning olulisusaste. Samuti valiti artikkel **I** tulemustele tuginedes instrumendi koostamiseks probleemi struktuur, milleks osutus strateegilise otsustamise problem. Lisaks on artiklis **II**avaldatud kirjandusanalüüsi alusel pakutud välja lahinguolukordades otsustamist mõõtvate mõõtevahendite liigutus: mõõtmine tegutsedes pärisolukordades, mõõtmine tegutsedes matkitud oludes, mõõtmine testimise teel. Artikkel **II** tulemusena valiti edasiseks tööks instrumendi tüüp, milleks osutus mõõtmine testimise teel, spetsiifiliseks testiks valiti olukorrapõhised otsustustestid (OPT).

Doktoritöö teise etapi peamiste tulemustena saab välja tuua, et artiklis **III** kirjeldatud protseduuridega valminud instrumendi kvaliteedinäitajad osutusid heaks ning test võimaldas edukalt eristada vastajaid väljaõppe taseme ja kogemuse alusel. Lisaks sellele näitas artiklis **III** kasutatud analüüs, et empiirikale tuginedes on võimalik instrumendi kvaliteeti veelgi tõsta, samuti on võimalik kasutada lisanduvaid andmeid uuteks kvaliteedi analüüsideks, mis omakorda võimaldab parandada instrumendi kvaliteeti veelgi. Seega selgus, et lisaks asjaolule, et instrumenti on võimalik kulutõhusalt ja paindlikult koostada, on selle kvaliteeti võimalik paindlikul moel parandada. Seega leiti, et koostatud instrument sobib madala juhtimistasandi sõjaväeliste juhtide otsustusoskuse mõõtmiseks lahingujuhtimise olukordades.

Doktoritöö kolmanda etapi tulemusena valminud artikli **IV** peamiste tulemustena võib esile tõsta asjaolu, et suurema kogemuse ja kõrgema väljaõppega kadettide testitulemused ületasid ajateenijate tulemusi. Seega leidis veelkord kinnitust, et väljaõpe ja varasem kogemus on olulised tegurid paremate tulemuste saamisel ja koostatud test sobib nende erinevuste hindamiseks. Küll aga selgus artiklis **IV**avaldatud uuringust, et stsenaariumi esitamise järgi (video ja paber) eristatud gruppide lahenduste kvaliteet statistiliselt olulisel määral ei erinenud. Samas selgus, et videoformaadis stsenaariumi tutvustuse saanud vastajad lahendavad teste kiiremini, kui need, kes said tutvustuse paberil.

Doktoritöö neljandal etapil valminud artikli **V** peamiste tulemustena võib esile tuua, et testide tulemuslikumat lahendamist prognoosivad eeskätt saavutatud väljaõppe tase (omandatud teamised ja oskused) ning selle käigus saadud kogemused, mida võib kokkuvõttes nimetada ekspertsuseks. Muud faktorid, mida selle PhD uuringu käigus loodud mudelis katsetati, ei avaldanud testi tulemuslikkusele märkimisäärset mõju vaatamata asjaoludele, et see baseerus teoreetilisele käsitusele ning mudeli kvaliteedinäitajad olid head.

Doktoritöö piirangutena esimeses etapis saab välja tuua asjaolu, et teadusajakirjades ei pruugi olla kajastatud kõik sobivad instrumendid, mida on kasutatud otsustamise mõõtmiseks lahingujuhtimise olukordades. Sedasorti uuringuid publitseeritakse palju spetsiifilistes raportites, misaga ei pruukinud olla kõik selle doktoritöö raames läbiviidud süstemaatilise kirjandusanalüüsiga leitavad. Piirangutena teises etapis saab välja tuua, et paljude asjaolude tõttu ei olnud võimalik läbi viia alternatiivseid instrumendi valiidsuse (ennustav valiidsus, korduv testimine jms) kontrolle. Samuti võib puudusena esile tõsta valimi väiksuse, mis on otseselt seotud ka eelmise puudusega, kuna valim oli kõikne ja suuremat valimit EKV tingimustes ühel ajahetkel pole võimalik komplekteerida. Seega polnud võimalik koguda andmeid, mis võimaldaksid kasutada erinevaid valideerimise meetodeid. Piirangutena kolmandas etapis saab välja tuua asjaolu, et testide lahendamiseks kulus kokku peaaegu kaks tundi, mis võis tekitada lahendajates tüdimust ning vähendada nende motivatsiooni. Ajateenijate testimiseks oli võimalik kasutada vaid õhtust aega, mis võis tähendada, et nad olid testimise ajaks juba päevastest tegevustest väsinud. Need asjaolud võisid mõjutada tulemusi. Oludest tulenevalt viidi testid läbi väikestes gruppides, mistõttu võis tekkida teatavaid erinevusi stsenaariumi esitamisel. Kadettide ja ajateenijate testimise ajal oli üks läbiviija vahetunud, ka see võis mõjutada stsenaariumite esitamist ja üldist õhkkonda testimisel. Seega ka need asjaolud võisid mõnevõrra mõjutada testi tulemusi. Neljanda etapi piirangutena võib esile tuua asjaolu, et vaatamata suurele hulgale andmetele sobitus vaid väike osa heade kvaliteedinäitajatega mudelise. Seega on põhjust arvata, et tuvastati vaid väike osa tulemuslikumat lahendamist prognoosivatest teguritest.

Doktoritöö tähtsusenateooria arendamisel on võimalikvälja tuua artiklis **II** pakutud mõõtevahendite klassifikatsiooni, samuti artiklis **III** välja töötatud instrumendi koostamise. Lisaks on artiklis **V** välja pakutud testide tulemuslikumat lahendamist prognoosiv mudel, mida saab kas kinnitada, ümber lükata või täiendada edasistes uuringutes. Lisaks kinnitavad selle töö tulemused objektiivse (väheste muude mõjudega) hindamise võimalikkust loodud teste kasutades. Samuti leidis kinnitust, et kasutades selle töö käigus väljatöötatud metoodikat, onvajadusel võimalik väheseid ressursse kaasates koostada mingite spetsiifiliste õpieesmerkide hindamiseks sobiv mõõtevahend.

Doktoritöö praktilise tähtsusena saab välja tuua artiklis **II** koostatud kirjandusülevaate, kuhu koondatut saavad kasutada edaspidi sõjandusvaldkonnas otsustamise uurimisest huvitatud teadlased või doktorandid, samuti ka muud huvilised. Artiklites **III** ja **IV** on detailselt kirjeldatud instrumendi koostamise protseduuri, mis võimaldab soovijatel seda korrata ja vajadusel edaspidi ka täiustada. Doktoritöö kaigus koostatud instrumenti saab EKV igapäevastes tegevustes kasutada mitmeti, näiteks sõjaväelise väljaõppe läbiviimisel, väljaõppe tulemuslikkuse hindamisel ning tõhustamisel. Lisaks sellele saab koostatud instrumenti kasutada KVA teadustöös, näiteks kadettide ja magistriõppe kuulajate lõputöödes. Instrument sobib lisaks olemasolevatele võimalustele kasutamiseks ka isikkooseisu valikute tegemisel ametikohtadele sobivuse määramisel, näiteks kandidaatide valimisel erinevatele kursustele.

10. ACKNOWLEDGEMENTS

First of all, I want to thank the University of Tartu, especially the Institute of Education. I really enjoyed my time in the nice, but old building at Salme 1a.

My special gratitude goes out to my supervisor Prof. Margus Pedaste. You were always friendly, did everything to encourage me and were patient with me when I struggled.

A very special thank you goes to my family – my wife Ave, son Markkus and daughter Margret. Thank you for supporting and always being with me!

I also want to express my gratitude to all cadets and conscripts who participated in my study. I enjoyed the time I spent with you during testing, and you were indeed devoted participants.

My special gratitude goes to Captain Robert Rajaste, Captain Meelis Sõukand and Siim Somelar. You were with me, as my co-fellows and helped me with many issues. I also want to thank my fellow doctoral students from the University of Tartu. I wish you all the best!

I especially want to acknowledge Roland Schimansky and Kairi Kasearu, who gave me a helping hand when I needed it, and also many thanks to Liina-Mai Tooding, for her assistance in my data analysis. Many thanks to other people in the Center of Excellence for Strategic Sustainability; the seminars I had with you were really interesting.

I also extend my warmest to the people of the Department of Applied Studies of Estonian Defence Academy. I spent five incredible years with you. Thank you for the support, unforgettable moments and being like a family to me.

I would not have started my studies without the unconditional support provided by Major General Martin Herem, who was the commander of the Estonian Defence College at that time. Thank you.

Finally, my gratitude is extended to the partners, who funded my PhD studies: the Estonian Ministry of Defence and the Estonian Defence Forces. It would not have possible without you.

11. PUBLICATIONS

CURRICULUM VITAE

Name: Tõnis Männiste
Date of birth: 02.07.1973
Citizenship: Estonian
Work address: Headquarters of the Estonian Defence Forces, Juhkentali 58, Tallinn 10132, Estonia
Phone: +372 522 7887
E-mail: tonis.manniste@mil.ee

Education

2015–	University of Tartu, Institute of Education, doctoral student
2007–2008	Baltic Defence College, Joint Command and General Staff Course
2007	Baltic Defence College, Army Intermediate Command and Staff Course
2003–2006	Tartu University, MA
1992–1996	Estonian Public Service Academy, BA

Employment

2019–	Headquarters of the Estonian Defence Forces, Training Department, Excercise Branch, Senior Staff Officer
2015–2019	Estonian National Defence College, Department of Applied Research, Resource management Group, junior research fellow
2014–2015	Estonian National Defence College, Training Department, Chief of Chair of Leadership and Pedagogy
2008–2014	Estonian National Defence College, Training Department, Chair of Tactics, Chief-Associate Professor of Military Pedagogics, Training Methodology Group, Military Pedagogy Field of Study
2005–2008	Estonian National Defence College, Chief of In-Service Centre
2003–2005	Estonian National Defence College, Chief of the Section of Electronic Study Aids of the Training Support Centre
2001–2003	Estonian National Defence College, Training Department, Chief of the Training Support Centre
2000–2001	Estonian National Defence College, Senior Instructor of Military Pedagogics and Physical Education
2000	Kuperjanov Infantry Training Bn, Commander of Staff Company
1998–2000	Kuperjanov Infantry Training Bn, BC Commander's Legal Adviser
1997–1998	Kuperjanov Infantry Training Bn, Commander of the Recce Platoon
1996–1997	Kuperjanov Infantry Training Bn, C Coy, Commander of the 1st Platoon
1996	Kuperjanov Infantry Training Bn, Training Coy, Commander of the 1st Training Platoon

Research interests
Testing in military sphere, especially decision-making and performance. Enhancing military training effectiveness.

Publications:

1. Kütt, K., Männiste, T. (2014). Kadettide arusaamad õppimisest ja õpetamisest. KVÜÕA Toimetised, 19, 103–126.
2. Männiste, T. & Pedaste, M. (2015). Probleemi mõiste sõjanduses ning probleemõppe lõimimisvõimalustest sõjaväelise väljaõppega. KVÜÕA toimetised, 20, 198–233.
3. Karton, I., Männiste, T., Tepp, A., Kornilov, I. (2018). Ajateenijate teenistusmotivatsiooni ja teenistusse jäämise soovi prognoosivad tegurid. Sõjateadlane, 6, 138–169.
4. Männiste, M., Rajaste, R., Suviste, R. & Pedaste, M. (2018). Olukorrapõhised otsustustestid rühmataseme sõjaväeliste juhtide otsustusvõime mõõtmiseks lahingujuhtimist matkivates olukordades. Sõjateadlane, 9, 63–118.
5. Männiste, T., Pedaste, M. & Schimanski, R. (2019). Review of Instruments Measuring Decision Making Performance in Military Tactical Level Battle Situation Context. Military Psychology, 31(5), 397–411. doi:10.1080/08995605.2019.1645538
6. Männiste, T., Pedaste, M. & Schimanski, R. (2019). Situational Judgement Test for Measuring Military Tactical Decision Making Skills. Military Psychology, *31*(6), 462–473. doi:10:1080/08995605.2019.1664366
7. Männiste, T., Pedaste, M., Kasearu, K. & Schimanski, R. (2019). Predicting Decision-Making Skills among Estonian Platoon-Level Military Commanders. *Military Psychology*, in the process of review. Submitted 15.08.2020.

ELULOOKIRJELDUS

Nimi:	Tõnis Männiste
Sünniaeg:	02.07.1973
Kodakondsus:	Eesti
Töökoha aadress:	Kaitseväe Peastaap, Juhkentali 58, Tallinn 10132, Eesti
Telefon:	+372 522 7887
E-post:	tonis.manniste@mil.ee

Hariduskäik

2015–	Tartu Ülikool, haridusteaduste instituudi haridusteaduse doktorant
2007–2008	Balti Kaitsekolledž, ühendstaabi ohvitseri kursus (Joint Command and General Staff Course)
2007	Balti Kaitsekolledž, jalaväebrigaadi staabiohvitseri kursus (Army Intermediate Command and Staff Course)
2003–2006	Tartu Ülikool, MA
1992–1996	Eesti Riigikaitse Akadeemia, BA

Teenistuskäik

2019–	Kaitseväe Peastaabi väljaõppeosakonna õppuste jaoskonna vanemstaabiohvitser
2015–2019	Kaitseväe Ühendatud Õppeasutuste rakendusuuringute keskuse teadus- ja arendustegevuse grupi nooremteadur
2014–2015	Kaitseväe Ühendatud Õppeasutused, Õppeosakond, Sõjaväe-pedagoogika ja juhtimise õppetooli ülem-lektor
2008–2014	Kaitseväe Ühendatud Õppeasutused, Õppeosakond, Taktika õppetooli sõjaväepedagoogika õppesuuna sõjaväelise välja-õppe metoodika grupi sõjaväepedagoogika suunaülem-dotsent
2005–2008	Kaitseväe Ühendatud Õppeasutused, Täiendõppe keskuse ülem
2003–2005	Kaitseväe Ühendatud Õppeasutuste väljaõppevahendite keskuse elektrooniliste õppevahendite sektsiooni ülem
2002–2003	Kaitseväe Ühendatud Õppeasutuste õppeosakonna planeerimis-jaoskonna õppevahendite keskuse ülem
2001–2002	Kaitseväe Ühendatud Õppeasutuste õppeosakonna õppe-vahendite jaoskonna ülem
2000–2001	Kaitseväe Ühendatud Õppeasutuste õppeosakonna sõjaväe-pedagoogika ja kehalise kasvatuse õppetooli vanemõppejõud
2000	Kuperjanovi üksik-jalaväepataljoni staabi- ja tagalakompanii ülem
1998–	2000 Kuperjanovi üksik-jalaväepataljoni pataljoniülema õigus-abi
1997–1998	Kuperjanovi üksik-jalaväepataljoni staabikompanii luurerühma ülem
1996–1997	Kuperjanovi üksik-jalaväepataljoni C-kompanii 1. rühma ülem
1996	Kuperjanovi üksik-jalaväepataljoni õppekompanii 1. rühma ülem

Teadustegevus
Testimine sõjanduses, otsustamise ja soorituse mõõtmine. Sõjaväelise väljaõppe tõhustamine.

Publikatsioonid
Vt inglisekeelne CV

DISSERTATIONES PEDAGOGICAE UNIVERSITATIS TARTUENSIS

1. **Карлеп, Карл**. Обоснование содержания и методики обучения родному языку во вспомогательной школе. Tartu, 1993.
2. **Ots, Loone**. Mitmekultuurilise hariduse õppekomplekt eesti kirjanduse näitel. Tartu, 1999.
3. **Hiie Asser**. Varajane osaline ja täielik keeleimmersioon Eesti muukeelse hariduse mudelitena. Tartu, 2003.
4. **Piret Luik**. Õpitarkvara efektiivsed karakteristikud elektrooniliste õpikute ja drillprogrammide korral. Tartu, 2004.
5. **Merike Kull**. Perceived general and mental health, their socio-economic correlates and relationships with physical activity in fertility-aged women in Estonia. Tartu, 2006.
6. **Merle Taimalu**. Children's fears and coping strategies: a comparative perspective. Tartu, 2007.
7. **Anita Kärner**. Supervision and research training within the professional research community: Seeking new challenges of doctoral education in Estonia. Tartu, 2009.
8. **Marika Padrik**. Word-formation skill in Estonian children with specific language impairment. Tartu, 2010.
9. **Krista Uibu**. Teachers' roles, instructional approaches and teaching practices in the social-cultural context. Tartu, 2010.
10. **Anu Palu**. Algklassiõpilaste matemaatikaalased teadmised, nende areng ja sellega seonduvad tegurid. Tartu, 2010.
11. **Mairi Männamaa**. Word guessing test as a measure of verbal ability. Use of the test in different contexts and groups. Tartu, 2010.
12. **Piret Soodla**. Picture-Elicited Narratives of Estonian Children at the Kindergarten-School Transition as a Measure of Language Competence. Tartu, 2011.
13. **Heiki Krips**. Õpetajate suhtlemiskompetentsus ja suhtlemisoskused. Tartu, 2011.
14. **Pille Häidkind**. Tests for assessing the child's school readiness and general development. Trial of the tests on the samples of pre-school children and first-grade students in Estonia. Tartu, 2011.
15. **Karmen Trasberg**. Keskkooli- ja gümnaasiumiõpetajate ettevalmistus Eesti Vabariigis (1918–1940) õpetajakoolituse ajaloolise kujunemise kontekstis. Tartu, 2011, 207 lk.
16. **Marvi Remmik**. Novice University Teachers' professional development and learning as a teacher: Opportunities and Conditions at Estonian Higher Education Institutions. Tartu, 2013, 129 p.
17. **Pilve Kängsepp**. Küsimuste kasutamine kui võimalus toetada õpilaste arusaamist loetust. Tartu, 2014, 125 p.

18. **Marge Täks**. Engineering students' experiences of entrepreneurship education. A qualitative approach. Tartu, 2015, 150 p.
19. **Reelika Suviste**. Students' mathematics knowledge and skills, and its relations with teachers' teaching and classroom management practices: Comparison between Estonian- and Russian-language schools. Tartu, 2015, 147 p.
20. **Liina Lepp**. The objectives of doctoral studies and factors influencing doctoral study process from the perspectives of different parties. Tartu, 2015, 271 p.
21. **Ülle Säälik**. Reading literacy performance: Metacognitive learning strategies matter, schools have effect on student outcomes. Tartu, 2016, 119 p.
22. **Katrin Saks**. Supporting Students' Self-Regulation and Language Learning Strategies in the Blended Course of Professional English. Tartu, 2016, 216 p.
23. **Anne Okas**. Novice and experienced teachers' practical knowledge in planning, delivery and reflection phases of teaching. Tartu, 2016, 172 p.
24. **Külli Kori**. The Role of Academic, Social and Professional Integration in Predicting Student Retention in Higher Education Information Technology Studies. Tartu, 2017, 168 p.
25. **Ingrid Koni**. The perception of issues related to instructional planning among novice and experienced teachers. Tartu, 2017, 142 p.
26. **Ivar Männamaa**. Development of an educational simulation game and evaluation of its impact on acculturation attitudes. Tartu, 2017, 154 p.
27. **Egle Säre**. Developing the reasoning skills of pre-schoolers through Philosophy for Children. Tartu, 2018, 131 p.
28. **Anu Sööt**. The procedure of guided core reflection for supporting the professional development of novice dance teachers. Tartu, 2018, 135 p.
29. **Tiina Anspal**. The development of teacher identity through role and self-conception in pre-service teacher education. Tartu, 2018, 157 p.
30. **Age Salo**. The dual role of teachers: school-based teacher educators' beliefs about teaching and understandings of supervising. Tartu, 2019, 156 p.
31. **Mirjam Burget**. Making sense of responsible research and innovation in science education. Tartu, 2019, 175p.
32. **Kaire Uiboleht**. The relationship between teaching-learning environments and undergraduate students' learning in higher education: A qualitative multi-case study. Tartu, 2019, 169 p.
33. **Karin Naruskov**. The Perception of Cyberbullying among Estonian Students According to Cyberbullying Types and Criteria. Tartu, 2020, 176 p.
34. **Raili Allas**. Supporting teachers' professional development through reflection procedure. Tartu, 2020, 182 p.
35. **Triinu Kärbla**. Assessment of text comprehension and teaching comprehension strategies in Estonian basic school. Tartu, 2020, 183 p.
36. **Kadi Luht-Kallas**. Risk-taking behaviour: Relationship with personality and markers of heritability, and an intervention to prevent unintentional injury. Tartu, 2020, 135 p.