DISSERTATIONES
PEDAGOGICAE
UNIVERSITATIS
TARTUENSIS
43

GERLI SILM

Test-Taking Motivation in Low-Stakes and High-Stakes Testing Contexts

DISSERTATIONES PEDAGOGICAE UNIVERSITATIS TARTUENSIS

43

DISSERTATIONES PEDAGOGICAE UNIVERSITATIS TARTUENSIS
43

GERLI SILM

Test-Taking Motivation in Low-Stakes and High-Stakes Testing Contexts

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 Education on June 17th 2022 by the joint Doctoral Committee of the Institute of Education and Institute of Ecology and Earth Sciences for awarding doctoral degrees in education, University of Tartu.

Supervisors: **Professor Margus Pedaste, PhD**
Institute of Education, University of Tartu, Estonia

Associate Professor emeritus Olev Must, PhD
Institute of Education, University of Tartu, Estonia

Associate Professor Karin Täht, PhD
Institute of Mathematics and Statistics, University of Tartu, Estonia

Opponent: **Associate Professor Hanna Eklöf, PhD**
Department of Applied Educational Science, Umeå University, Sweden

Commencement: The White Hall of The University of Tartu Museum, Lossi 25, Tartu, on September 6th, 2022, at 10 a.m.

The research has been supported by the European Regional Development Fund measure “Increasing the local socioeconomic impact of the R&D system and smart specialization for the development of growth areas” activity 4.2.2 “Higher education specialist scholarships in the growth areas of smart specialization”, by the European Social Fund (DoRa and Doctoral School), and by the University of Tartu ASTRA Project PER ASPERA, financed by the European Regional Development Fund.

European Union
European Regional
Development Fund

Investing
in your future

ISSN 1406-1317
ISBN 978-9949-03-983-8 (print)
ISBN 978-9949-03-984-5 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF FIGURES 7

LIST OF TABLES 7

LIST OF ABBREVIATIONS 8

LIST OF ORIGINAL PUBLICATIONS 9

1. INTRODUCTION 10
   - 1.1. Research Problem 10
   - 1.2. Focus of the Current Research 13
   - 1.3. Terminology Used in the Dissertation 15

2. THEORETICAL FRAMEWORK FOR TEST-TAKING MOTIVATION 17
   - 2.1. Validity of Test Score Interpretation 17
   - 2.2. The Role of Motivation in Achievement Behaviour 18
   - 2.3. Test-Taking Motivation 21
   - 2.4. Assessment of TTM 22
     - 2.4.1. Self-Report Measures of TTM 22
     - 2.4.2. Measures of TTM Based on Test-Taking Time 23

3. RESEARCH DESIGN AND METHODS 24
   - 3.1. Educational Testing in Estonia 24
   - 3.2. Research Design 25
   - 3.3. Procedure and Participants 25
   - 3.4. Measures 26
   - 3.5. Data Analysis 28

4. FINDINGS 29
   - 4.1. In the LS Context SRE and RTE Complemented Each Other and Predicted a Significant Part of the Variance in the Test Results 29
   - 4.2. Test-Taking Effort had a Significant Relationship with Test Performance in the HS Context 32
   - 4.3. The Average Meta-Analytic Correlation Between Test-Taking Effort and Performance was Significantly Different When Measured Using SRE or RTE 33

5. DISCUSSION 35

6. LIMITATIONS 37

7. CONCLUSIONS AND IMPLICATIONS 38
   - 7.1. Theoretical Relevance 38
   - 7.2. Practical Relevance 38

REFERENCES 40
SUMMARY IN ESTONIAN 46
ACKNOWLEDGEMENTS 49
PUBLICATIONS 51
CURRICULUM VITAE 150
ELULOOKIRJELDUS 152

LIST OF FIGURES

Figure 1. Eccles Expectancy Value Model of Achievement Choices (from Eccles & Wigfield, 2020) 20
Figure 2. Overview of Research Design 25
Figure 3. Model Predicting Test Performance in a LS Testing Context with Number of Answered Items and Average Time per Items... 29
Figure 4. Model Predicting Test Performance in a LS Testing Context with SRE and RTE 30
Figure 5. Model Predicting Test Performance in a LS Testing Context with SRE and Average Time Spent on Incorrect Item 31
Figure 6. Model Predicting Test Performance in a HS Testing Context with SRE and RTE 32
Figure 7. Meta-Analytic Correlations Between Test-Taking Effort and Test Performance 34

LIST OF TABLES

Table 1. Sample Description for the LS and HS Studies 26

LIST OF ABBREVIATIONS

CIV	–	construct irrelevant variance
EV theory	–	expectancy-value theory
HS	–	high-stakes
IRT	–	item response theory
LS	–	low-stakes
PISA	–	Programme for International Student Assessment
RTE	–	Response Time Effort
SEM	–	Structural Equation Modelling
SEVT	–	Situated Expectancy-Value Theory
SOS	–	Student Opinion Scale
SRE	–	self-reported effort
TTM	–	test-taking motivation

LIST OF ORIGINAL PUBLICATIONS

The dissertation is based on the following original publications, which are referenced in the text by their Roman numerals. The publications are listed in the chronological order of publication.

I. Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *TRAMES: A Journal of the Humanities & Social Sciences*, *17*(4), 433–448.

II. Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. *TRAMES: A Journal of the Humanities & Social Sciences*, *23*(3), 353–376.

III. Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*, 100335.

IV. Silm, G., Must, O., Täht, K., & Pedaste, M. (2021). Does test-taking motivation predict test results in a high-stakes testing context? *Educational Research and Evaluation*, 1–27.

Author contributions:
The current dissertation integrates three studies: two empirical studies and a systematic literature review with a meta-analysis. The first empirical study is part of a wider study ("Factors influencing the academic performance in university students" within PRIMUS programme) looking at first year students in higher education. In this study the author's role was related to measuring cognitive abilities, i.e. compiling the instrument and analysing the data. Articles I and II are based on this study. The data for the second empirical study came from a real-life setting of university admission testing. This study is the basis for Article IV. The literature review and meta-analysis is presented in Article III. In all of the articles the author carried the role of lead author, formulated research questions, planned the data collection, and analysed the data in co-operation with the co-authors.

1. INTRODUCTION

1.1. Research Problem

Testing in education and psychology has a long history. There is evidence of strict standardised examinations from ancient China dating back to 2200 B.C.E (Geisinger & Usher-Tate, 2016). Test results have informed important decisions, for example about soldier's position and capability of serving in the army (Army Alpha and Army Beta tests used during World War I; Yoakum & Yerkes, 1920) or identifying intellectually disabled children (Binet & Simon, 1905, as cited in Geisinger & Usher-Tate, 2016). Testing has quite a long history in Estonia as well: the first well-known large-scale standardised testing comes from the 1930s when Juhan Tork carried out intelligence testing among Estonian children (Tork, 1940).

In the current research and practice of education and psychology, tests are extensively used to evaluate and assess people's various characteristics and abilities. Students come across several tests in their normal day to day school life. These can be in the form of in-class tests and regular exams, but also standard-determining tests, admission tests, international tests, etc. There are possibly many instances where the test result does not bring a personal consequence to the test-taker. For example, there are national or international studies and surveys that do not provide personal feedback to the participants. Based on the received test results inferences are made about certain people or groups of people. For example, results from international tests can form our perception about the differences between students and teachers in different countries. On a macro level, test results can be basis for different political decisions and educational strategies.

It is common in the field of education that results from certain tests are used to evaluate the quality of schools and teachers, resulting in rankings and league tables. For instance, in Estonia every year ranking of schools based on their students' average state exam results are published. Many believe that the position in the ranking indicates the quality of schools and its teachers, and often alternative explanations and the statistical soundness are overlooked (e.g., Goldstein, 2004, 2014). Results from the PISA tests shape the understanding about entire education systems. But can there be more to these results?

The problem with educational achievement tests as stated by Messick (1984, p. 216 as cited in Haladyna & Downing, 2004) is that the tests "at best, reflect not only the psychological constructs of knowledge and skills that are intended to be measured, but invariably a number of contaminants". These contaminants include various psychological and situational factors that constitute as construct irrelevant variance (CIV) in test scores (Messick, 1984, p. 216 as cited in Haladyna & Downing, 2004), which in turn is a threat to the validity of test score interpretation. Possible sources of CIV include: reading comprehension, test anxiety, (test) fatigue, motivation to perform on a test, item formats, item quality, differential item functioning, test administration conditions, test preparation, computer-based testing, and others (Haladyna & Downing, 2004). Being not aware of CIV in test results is a threat to the validity of inferences based on test results.

The role of motivation in testing has been acknowledged long time ago. The famous educational psychologist Cronbach (1990, p. 79) has stated that "unless he (examinee) cares about the result, he cannot be measured" indicating that personal motivation to take a test is a relevant factor for obtaining valid results. The important role of motivation in performing educational tasks is also supported by contemporary theories of human motivation. When earlier theories of human behaviour relied mainly on the relationship between stimuli and response, contemporary theories of motivation take into account the internal processes within individuals (Schunk et al., 2014). This means that the same stimuli can produce different outcomes for different persons depending on their prior experiences, attitudes, and other internal processes. It is important to take this into account in the context of testing. When students come across many different testing situations, their motivation to perform at their best may vary and this can bring CIV to test results.

In the last decades, test-taking motivation (TTM) as a possible source of CIV has received the attention of several researchers. The importance of studying TTM has stemmed from the fact that many tests do not have personal consequences for the test-takers. In other words, the tests for them individually are low-stakes (LS), even if the tests might be high-stakes (HS) on some other levels (e.g., institutional or national level). In these cases the test-takers may not be motivated to exert maximum effort when taking the test and therefore do not show their actual abilities, knowledge, or skills. It has been found that in low-stakes testing contexts test results are on average lower than in high-stakes testing contexts (DeMars, 2000; Duckworth et al., 2011; Napoli & Raymond, 2004; Sundre, 1999; Wolf & Smith, 1995). In their meta-analysis Wise and DeMars (2005) found the difference in performance between LS and HS testing contexts to be 0.59 standard deviations.

In HS testing contexts, TTM has been studied to a lesser extent, because performance in these tests is considered to be in the interests of the test-taker (Wise, 2020). In reality, it is not always easy to distinguish between LS and HS testing contexts. According to expectancy-value (EV) theory (Eccles & Wigfield, 2002; Eccles & Wigfield, 2020; Rosenzweig et al., 2019; Wigield & Eccles, 2000), motivation on a task depends on the individual's perception of their own competence and value of the task. Therefore, the stakes of taking a certain test vary individually. There can be occasions of seemingly high-stakes testing situations where TTM can be an important aspect to take into account for a more valid interpretation of test results.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p.11) state that "it is important to determine whether test takers regard the test experience seriously, particularly when individual scores are not reported to test takers or when the scores are not associated with consequences for the test takers. Decision criteria regarding whether to include scores from individuals with questionable motivation should be clearly documented" (AERA, APA, & NCME, 2014, p. 213). The need for documenting the motivation of test-takers has been acknowledged for example in the PISA study, which is the world's largest study of pupils' scholastic performance. Therefore, measures of TTM such as effort thermometer (Kunter et al., 2005) and Response Time Effort (RTE; Wise & Kong, 2005) have been included in the PISA studies (OECD, 2019b). There are studies suggesting that students' effort declines during the test (Pools & Monseur, 2021), which can have an impact on countries' rankings based on the average PISA scores (Akyol et al., 2021). Effort has been shown to explain as much as 32–38% of the variance in PISA scores across countries, but not within countries (Zamarro et al., 2019).

In order to consider the possible role of TTM in test results, TTM needs to be assessed. So far TTM has been measured mainly with self-report instruments (e.g., Student Opinion Scale (SOS), Sundre & Moore, 2002). However, it is known that self-reports can be subject to various biases, such as social desirability in answering. Because of this, time-based measures, such as RTE (Wise & Kong, 2005) have been offered as an objective, non-obtrusive alternative to self-report measures. The use of time-based indicators of TTM is possible due to the use of computer-based tests that allows monitoring test-taking time on item level.

More and more tests are being transferred from paper-pencil format to Internet- and computer-based formats. Also, the PISA study has to a large extent become computer-based: in 2018 most of the test-takers took the computer-based test and only nine countries used the paper-pencil test (OECD, 2019a). In Estonia, standard-determining tests, national examinations, and university admission tests are too becoming computer-based (Eurydice, 2020b; Johanson et al., 2021; Puksand, 2017; Tartu Ülikool, 2022). On one hand, computer-based testing brings more unknown elements to testing – pupils may perceive and fill these differently compared to traditional paper-and-pencil tests (e.g., Clariana & Wallace, 2002; Kolen & Brennan, 2004, pp. 316–320). On the other hand, computer-based testing brings more opportunities to observe and analyse test-taking behaviour and (e.g., Wise & Kong, 2005). However, with the objective, time-based indicators of effort there is the question of whether and how much the observation of behaviour reflects the underlying mental mechanisms such as motivation.

Although sometimes the terms TTM and test-taking effort have been used as synonyms (Wise & Gao, 2017), in this dissertation I will make a distinction between the two. The connection between the two can be seen as described by Lundgren and Eklöf (2020): "motivation will regulate the maximum amount of effort they (test-takers) are willing to spend on pursuing the test-taking goal." It means that TTM and test-taking effort are not always equal – sometimes maximal effort is not necessary for obtaining the desired result. When the task is relatively easy for the test-taker, but if the test-taker is motivated to take they will exert the necessary amount of effort. Several studies have shown that compared to other components of TTM, test-taking effort is the best predictor of test performance (Cole et al., 2008; Knekta & Eklöf, 2015; Penk & Schipolowski, 2015). Also, when using time-based measures as indicators of TTM, they are rather indicators of test-taking effort in specific, not TTM in general, because they reflect the behaviour and not the internal motivational processes of test-takers. In this dissertation I will concentrate on test-taking effort as one component of TTM.

Knowledge about individual test-takers' TTM or test-taking effort gives an opportunity to choose whether or not to include the scores of test-takers with low motivation to the overall test results. The procedure of not including these scores is called motivation filtering. It has been shown that test results differ significantly depending on whether or not data from unmotivated test-takers is included (Wise & DeMars, 2010; Wise et al., 2006).

With growing need to take TTM into account especially in LS tests, it becomes increasingly important to know more about the approaches that can be used to measure TTM and test-taking effort. For example, there is at least one study claiming that self-reported effort (SRE) and RTE are equally good for filtering out the data from test-takers with low TTM (Swerdzewski et al., 2011). However, other studies show that the correlation between the SRE and RTE has remained low to moderate (Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005), indicating that the two methods reflect different aspects of test-taking effort. The difference between objective and subjective effort has been observed in other contexts as well (e.g., Apascaritei et al., 2021). This raises the question: what aspects of TTM do self-reported effort and time-based measures of effort reflect and how are these measures applicable in different test-taking situations.

The main aim of the current dissertation is to study the possible effect of test-taking effort on test performance in both LS and HS testing contexts, using two approaches (self-reports and behavioural estimates based on the use of test-taking time) and comparatively evaluating the obtained test-taking effort estimates in different contexts. This provides a more comprehensive view of measuring TTM and test-taking effort, and guides practitioners who need to consider, measure and report the level of test-takers' motivation and effort for more valid interpretation of test results. To fill this aim three studies were undertaken. I will describe the three studies and their results in Chapters 3 and 4 respectively, after an overview of the concepts and theories necessary to understand the wider framework in which TTM belongs in Chapter 2.

1.2. Focus of the Current Research

Proceeding from previous findings about TTM and test-taking effort, and the theoretical framework for TTM, assumptions about the possible findings were made. In the context of TTM it is important to take into account test-takers' cognitive abilities, to confirm that the measured TTM is not just a proxy of cognitive abilities (Gagné & St Père, 2001; Penk & Richter, 2017; Reeve & Lam, 2007; Wise & DeMars, 2005). It has been shown that performance in educational assessments and cognitive ability tests have a large common variance (Deary et al., 2007; Neisser et al., 1996); accordingly, the best predictor of performance is previous performance on a similar task (Goldstein, 1997). Concluding from this, results from national examinations (previous performance on a HS test) has been used as an indicator of cognitive abilities in the dissertation. Also, many previous studies have shown the relationship between gender and educational performance, although the underlying reasons are not always clear (Coluccia & Louse, 2004; Hyde et al., 1990; Hyde & Linn, 1988; Linn & Hyde, 1989; Núñez-Peña et al., 2016; Voyer & Voyer, 2014). It has been proposed that these differences stem from differences in motivation (Pekkarinen, 2015; Steinmayr & Spinath, 2008). Therefore, both of these – previous performance and gender – were considered important control variables. Based on this the first hypothesis is proposed:

1) **In a LS testing context test-taking effort is related to test performance even when previous performance and gender are controlled for.**

Most of the studies on TTM have been conducted in LS contexts, as in such cases low motivation is seen as a serious problem. Studies on TTM in HS testing contexts are few, and the results have been contradictory. Sundre and Kitsantas (2004) found no effect of TTM on test results in a HS context in their experimental study. On the other hand, Knekta (2017), Knekta and Eklöf (2015), and Stenlund et al. (2017) observed significant positive relationships between TTM and test performance in real-life HS testing contexts. Knekta and Eklöf (2015) find it relevant to study TTM in HS contexts to increase knowledge about students' perception of tests in a wider psycho-educational sense. In addition, data from HS testing contexts can be used for other purposes, for example item calibration. In such cases it would be important to know whether TTM has influenced the test results. Also, the distinction between LS and HS tests is not always clear; there can be situations that lie in between the two extremities. Drawing on EV theory, motivation depends on the inner processes and previous experiences of the test-taker, inferring that the stakes of a testing situation can be perceived very differently depending on the individual. Hence, the second hypothesis is:

2) **In a testing context that can be considered HS test-taking effort is related to test performance when previous performance and gender are controlled for.**

Wise and Kong (2005) have proposed the Response Time Effort (RTE) as an alternative to self-report measures of TTM. However, RTE is shown to have ceiling effect (Wise & Kong, 2005), i.e. for most examinees the test-taking effort is considered to be the highest value. Swerdzewski et al. (2011) have concluded that RTE and SRE are both able to filter out data of unmotivated examinees from the entire dataset. Nevertheless, the relationship between SRE and RTE has been shown to be low to moderate (Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005). This leads to the third hypothesis:

3) **Self-reported effort and time-based measures of effort complement each other, sharing some common variance but also predicting test performance independently.**

A more general overview of different methods used for assessing test-taking effort was obtained with the literature review. Within this a meta-analysis was conducted to find the meta-analytic average correlations between test performance and test-taking effort based on time-based and self-reported indicators of effort (SLR study). Testing the hypotheses in empirical studies involved gathering data about test-taking effort (both self-reported and time-based) in LS and HS testing contexts, and creating structural equation models (SEM) with effort indicators predicting test performance when previous performance and gender are controlled for (LS and HS study). This enabled determining how these predict performance separately and in combination and what is their relation to other variables in the model.

1.3. Terminology Used in the Dissertation

In this section, I will introduce and explain the main terms used throughout the dissertation. First, **measurement** means the assignment of scores to individuals with the scores representing a certain characteristic of the individuals (Price, 2012). The characteristics measured within this dissertation such as cognitive ability and test-taking effort are not directly observable but latent. This leads to another important term, which is construct. **Construct** refers to the concept or characteristic a test is designed to measure (AERA, APA, & NCME, 2014, p. 11). Haladyna and Downing (2004) stress that defining the construct is the most fundamental step in validation. Cognitive ability and test-taking effort mentioned above are examples of such constructs.

The aim of testing is to find out how examinees differ from each other in terms of the measured construct. Haladyna and Downing (2004) emphasize two achievement constructs that are relevant in educational and psychological testing: a domain of knowledge and skills (declarative or procedural knowledge), and cognitive ability (reading, writing, mathematical problem-solving). When talking about **achievement tests**, I mean tests that measure either of these, but in the empirical studies of this dissertation performance on a cognitive ability test is in focus.

The concept of **validity** is essential to this dissertation. Simply said, by validity it is meant whether it is assessed what is intended. In this thesis it is important in two aspects. First, TTM is an issue related to validity, because it may bring unwanted variance to assessment test results. For example, the results from a cognitive ability test taken in a LS testing situation may reflect not only cognitive abilities, but also TTM. If this is not taken into account, the interpretations based on the test results may not be valid. Second, when measuring TTM we want to be sure that the instruments used measure TTM and are not a reflection of for example self-serving bias in self-report questionnaires. Important aspects of validity are further discussed in Chapter 2.1.

In terms of validity it is important to consider the **test-taking context**. When talking about testing situations or contexts, I mean any kind of assessment situation including exams, in-class tests, (international) studies of abilities, skills, competencies, etc. In TTM studies a distinction is made between **low-stakes (LS) and high-stakes (HS) tests** or, more specifically, testing contexts depending on how relevant the test is for the test-takers (e.g., Cole & Osterlind, 2008; Mislevy, 1995; Sundre & Kitsantas, 2004). However, it has to be noted that talking about LS and HS tests is not entirely correct. A test itself cannot be LS or HS; the way the results of the test are used is what makes a testing context LS or HS. It can also be argued that the stakes of the test are determined by the individual taking the test. According to the EV theory of motivation, the same test in the same testing context could be personally relevant for one individual and non-relevant for another depending on various personal characteristics such as self-schemata, motives for taking the test, previous experiences, self-efficacy, etc.

In the current thesis when referring to LS or HS tests or testing contexts/ situations, I mean that a LS **testing context is non-consequential**, i.e. there is no relevant observable outcome for the test-taker. This, however, does not exclude a possibility that some test-takers consider these tests personally relevant for comparative, competitive, or other reasons, e.g. feeling of social responsibility (Eklöf, 2006). A **HS testing context on the other hand is consequential**, i.e. taking the test has a direct consequence for the test-taker. Again, it has to be recognized that there are test-takers who do not care about for example grades and therefore do not consider the test situation as HS for them. For example, for most test-takers national examinations are probably HS as they can be an important prerequisite for continuing into higher education, whereas for some who do not wish to go to university or are accepted on different grounds, national examinations may not be personally relevant. Other examples of HS situations include graded school assignments, exams, and admission exams. Results from the latter potentially impact a person's educational path and career choice, and can therefore be related to income, welfare, etc. in later life.

2. THEORETICAL FRAMEWORK FOR TEST-TAKING MOTIVATION

2.1. Validity of Test Score Interpretation

According to the *Standards for Educational and Psychological Testing* (hereafter the Standards; AERA, APA, & NCME, 2014, p.11), validity "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of the test". Similarly, Messick (1993, p.13) has defined validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment." He stresses that not a test or observation device as such needs to be validated but the inferences drawn from these (p.13). This means that "tests do not have reliabilities and validities, only test responses do" and test responses "are a function not only of the items, tasks, or stimulus conditions, but the *persons* responding and the *context* of measurement" (p.14). Messick (1993, p.13) also stresses that validity is "a matter of degree, not all or none" and that "validation is a continuing process" meaning that interpretations of validity can change in light of new information.

As advocated by Messick (1993) the latest versions of the Standards (AERA, APA, & NCME, 1999, 2014) do not refer to distinct types of validity (content validity, criterion-related validity, predictive validity, concurrent validity, construct validity) as has been the tradition since at least the early 1950s (Messick, 1993, p.16) but see validity as a unitary concept. Instead, the Standards (2014, pp. 26–31) refer to different types of validity evidence: content-oriented evidence; evidence regarding cognitive processes; evidence regarding internal structure; evidence regarding relationships with conceptually related constructs; evidence regarding relationships with criteria; evidence based on consequences of tests.

For valid interpretations of test scores it is important to specify the construct the test intends to measure (e.g., general ability, depression, self-regulation etc.). A uniform interpretation of a test score is considered rare. For example, using the same test in different settings implies a somewhat different interpretation of the test scores. There can be cases when a test measures less (construct underrepresentation or construct deficiency) or more (construct irrelevant variance or construct contamination) than its proposed construct (AERA, APA, & NCME, 2014).

The Standards emphasise that motivation of the test-takers should also be considered for more valid interpretation of test results. More specifically, it is indicated that there are several conditions related to test administration that might affect item performance. These may include (AERA, APA, & NCME, 2014, p. 89) "motivation of the test takers, item position, time limits, length of test, mode of testing (e.g., paper-and-pencil versus computer administered), and use of calculators or other tools." Low levels of effort and motivation may result in inappropriate score interpretations. The Standards (AERA, APA, & NCME, 2014, p. 213) consider the degree of motivation to be a type of information "relevant to the interpretation of test results in policy settings."

2.2. The Role of Motivation in Achievement Behaviour

The term "motivation" is derived from the Latin word "*movere*", which means "to move" (Schunk et al., 2014). The idea of movement is also present in the definition of motivation by Schunk, Pintrich, and Meece (2008, as cited in Ryan, 2012, p. 13): "motivation is the process whereby goal-directed activities are energized, directed and sustained." Motivation is a process that cannot be directly observed but rather its presence can be inferred from actions such as choice of tasks, effort, persistence, and verbalizations (e.g., "I really want to work on this") (Schunk et al., 2014, p. 5). Beside physical actions, motivation requires mental actions such as "planning, rehearsing, organizing, monitoring, making decisions, solving problems and assessing progress" (Schunk et al., 2014, p. 5). All of these actions are essential in education and moving towards goals.

Early views of human behaviour postulated that differences in motivation were related to individual differences in instincts and traits, or level of responding to stimuli caused by reinforcements or rewards (Ryan, 2012; Schunk et al., 2014). Contemporary theories of cognitive motivation hold that individuals' thoughts, beliefs, and emotions underlie motivation (Ryan, 2012). For example, social cognitive theory of motivation postulates that people's actions are in accordance with beliefs about their capabilities and expected outcomes of the actions (Ryan, 2012). Many motivational theories have historically included the components of expectancies and values (e.g., Atkinson, 1957; Schunk et al., 2014).

Expectancies refer to an individual's beliefs about their competence and capability to perform a certain task successfully. If expected to fail or having experienced failures, most individuals will eventually not choose to work or give effort in the task even if the task has value for them (Schunk et al., 2014). Values are individuals' beliefs about the reasons for engaging in certain tasks. There may be a variety of reasons for wanting to perform a task: it may be viewed as interesting, enjoyable, important or useful, it may provide a reward or help to avoid punishment, etc. (Schunk et al., 2014).

A contemporary EV theory relevant to educational research has been developed by Eccles, Wigfield, and their colleagues (e.g. Eccles & Wigfield, 2002, Wigfield & Eccles, 2000; Schunk et al., 2014). The latest version of the theory is called situated expectancy-value theory (SEVT) (Eccles & Wigfield, 2020). In this theory, the two most important predictors of achievement behaviour are *expectation of success* and *subjective task value.* Expectation of success refers to the individual's belief about how well they will do on the upcoming task (Eccles & Wigfield, 2020). This is somewhat similar to the concept of self-efficacy by Bandura (1977). Subjective task value consists of interest-enjoyment value, attainment value (importance of the task), utility value (usefulness of the task), and relative cost. Intrinsic value or interest value is the anticipated enjoyment expected from doing the task. Utility value or usefulness is conceptualised as how well doing the task fits in the person's present or future plans. Attainment value is the relative identity-based importance of engaging in certain tasks or activities. Perceived cost refers to the idea that every activity is related to some cost, and individuals tend to avoid the situations where the cost overweights the benefits. There are different types of costs: 1) effort cost (the perception of the effort needed to complete the task and whether it is worth it); 2) opportunity cost (the extent to which doing one task inhibits doing other valued tasks); 3) emotional costs (e.g. anticipated anxiety, feelings related to failure) (Eccles & Wigfield, 2020).

Subjective task value and expectancy of success are just one part of the SEVT model concentrating on aspects of individual decision making. The formulation of these is related to individual's various previous experiences. These are illustrated in Figure 1. The middle part of the model represents developmental processes related to development of self-concepts and memories, and the left side of the model focuses on the world in which individuals mature, including cultural context (Eccles & Wigfield, 2020; Rosenzweig et al., 2019).

TTM is often conceptualised in the expectancy-value framework (Silm et al., 2020). If an individual is asked to take a test then, according to SEVT, they will consciously or unconsciously consider the following aspects when deciding whether and how much effort should be spent on the test:

- How I think I will do on the test, am I good at tasks like this, how I have done previously on similar tasks? (expectancy of success);
- Do I expect the taking the test to be interesting or enjoyable? (interest value);
- Will taking the test be consistent with my self-concept (e.g. as a good student I will give my maximum effort on the tasks I am asked to do)? (attainment value);
- Will taking the test be useful for me (e.g. enables admission to university, a good grade earns parental approval etc.)? (utility value);
- What are the possible costs related to taking the test, e.g. how much effort the test will take? (effort cost);
- What else could I be doing instead of taking the test? (opportunity cost);
- What emotions could accompany test-taking, e.g. anxiety, fatigue, fear of failure, threat to self-concept? (emotional cost).

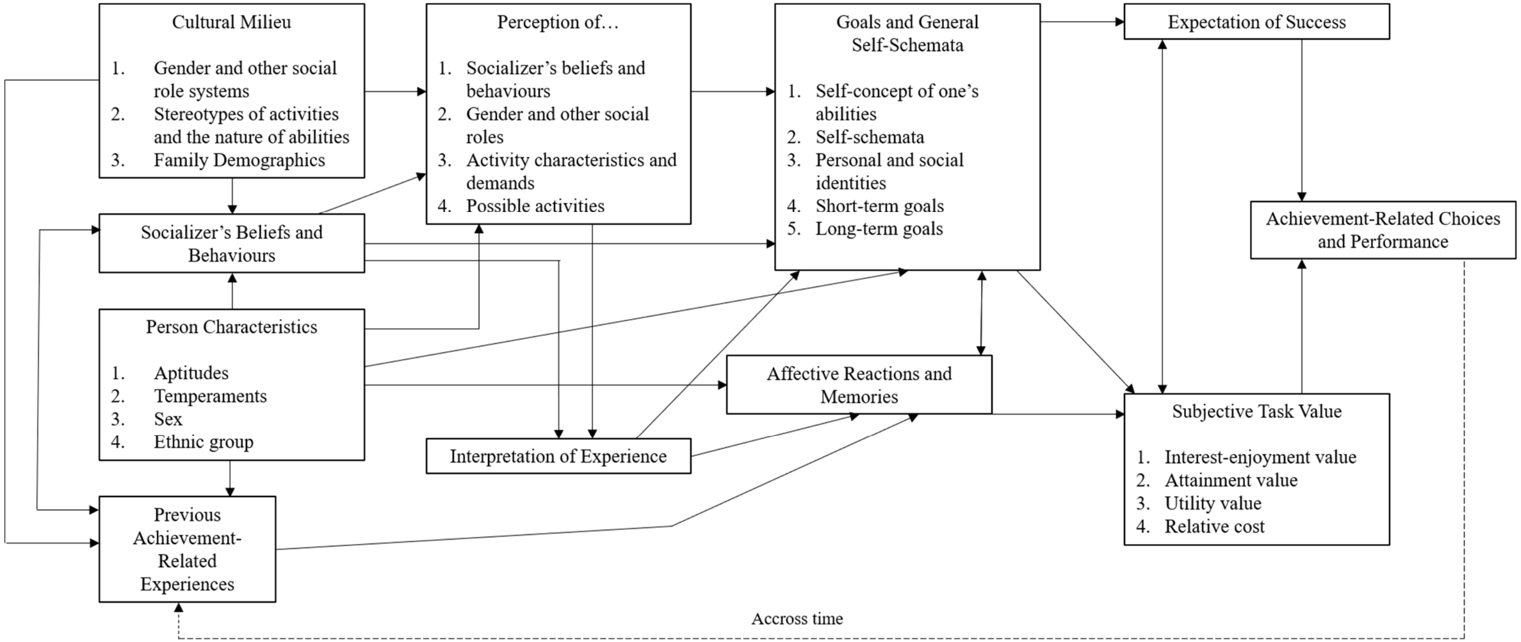

Cultural Milieu
1. Gender and other social role systems
2. Stereotypes of activities and the nature of abilities
3. Family Demographics
Perception of…
1. Socializer's beliefs and behaviours
2. Gender and other social roles
3. Activity characteristics and demands
4. Possible activities
Goals and General Self-Schemata
1. Self-concept of one's abilities
2. Self-schemata
3. Personal and social identities
4. Short-term goals
5. Long-term goals
Expectation of Success
Achievement-Related Choices and Performance
Socializer's Beliefs and Behaviours
Person Characteristics
1. Aptitudes
2. Temperaments
3. Sex
4. Ethnic group
Affective Reactions and Memories
Interpretation of Experience
Subjective Task Value
1. Interest-enjoyment value
2. Attainment value
3. Utility value
4. Relative cost
Previous Achievement-Related Experiences
Accross time

Figure 1. *Eccles Expectancy Value Model of Achievement Choices (from Eccles & Wigfield, 2020).*

2.3. Test-Taking Motivation

The concept of test-taking motivation is a relatively new one, as it emerged a few decades ago, first in the context of personnel selection (Arvey et al., 1990). Nevertheless, motivation has been seen as a relevant issue in test taking for much longer, almost since the first psychological and educational tests, such as Binet's IQ test in 1905. For example, wrong answers in tests have been penalized with the aim to reduce guessing behaviour that may result from low motivation for taking the test (e.g., Holzinger, 1924; Thurstone, 1919). Spearman (1927) found that effort can influence intelligence tests, referring to Webb who suggested that there is a *w* (will) factor in addition to the *g* factor (general intelligence). Cronbach (1990, p. 79) also stressed the relevance of motivation and stated that if the examinee does not care about the results, they cannot be measured.

In recent decades a more specific term of test-taking motivation (TTM) has come into use. TTM has been defined as "giving one's best effort to the test, with the goal being to accurately represent what one knows and can do in the content area covered by the test" (Wise & DeMars, 2005, p. 2). Knekta and Eklöf (2015, p. 662) have specified that it is "a situation-specific motivational construct". Cheng (2014, p. 306) has concluded that "motivation varies according to the complex interaction of test-takers and test contexts based on both the intended and unintended test use." Another relevant term is test-taking effort, that is test-taker's "engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (Wise & DeMars, 2005, p. 2). It has been shown that from the different aspects of motivation, test-taking effort is the best predictor of performance (Cole et al., 2008; Finney et al., 2018; Penk & Schipolowski, 2015; Zilberberg et al., 2014). In the context of EV theory effort occurs when the person is motivated, so effort can be seen as a reflection of motivation in behaviour. Motivated students are more likely to expend greater mental effort and employ better strategies (Schunk et al., 2014).

TTM has been mostly viewed in the context of EV theory by Eccles and Wigfield (2000, 2002). It can be said that TTM consists of three components: test-taking effort, expectancy of success, and value of the test. Test-taking effort stems from the outcome of expectancy and value (Penk & Schipolowski, 2015; Wigfield & Eccles, 2000; Ulitzsch et al., 2021).

It has been shown that the results of HS tests are approximately 0.5 standard deviations higher than the results of LS tests (Wise & DeMars, 2005). Results from motivation filtering studies show that the average performance improves when the results from persons with low test-taking motivation are filtered out (Wise & DeMars, 2010; Wise et al., 2006). Motivation filtering is a procedure during which the data of apparently unmotivated test-takers are removed from the dataset. In order to distinguish unmotivated test-takers from motivated test-takers, and achieve a more valid interpretation of test results, it is necessary to measure TTM.

2.4. Assessment of TTM

Some commonly employed indexes of motivation are: choice of tasks, effort, persistence, and achievement. Motivation can be estimated through direct observations, ratings by others, and self-reports. Assessing motivation is most often done by using self-reports (Schunk et al., 2014). This also applies to TTM (Silm et al., 2020). Self-reports "capture people's judgments and statements about themselves" (Schunk et al., 2014).

Direct observations of motivated behaviour (choice of tasks, effort expended, and persistence) are valuable indicators of motivation because in these the inferences of the observer plays only a minor role (Schunk et al., 2014). Logging and analysing test-taking times is also a direct observation. The critique of direct observations is that observations "ignore the cognitive and affective processes underlying motivated behaviours" (Schunk et al., 2014).

2.4.1. Self-Report Measures of TTM

As stated above, TTM is usually estimated with self-reported questionnaires presented after the test. The most frequently used questionnaire is the Student Opinion Scale (Sundre & Moore, 2002; Silm et al., 2020). But there are others, for example Questionnaire of Current Motivation (Rheinberg, Vollmeyer, & Burns, 2001), Test-Taking Motivation Questionnaire (Eklöf, 2010), Valence, Instrumentality, Expectancy Motivation Scale (Sanchez, Truxillo, & Bauer, 2000), PISA effort thermometer (Kunter et al., 2002), Online Motivation Questionnaire (Boekaerts, 2002), Motivation Questionnaire (Knekta & Eklöf (2015). Knekta and Eklöf (2015) noticed that many of the questionnaires take into account only a selection of aspects of the EV theory. Their Motivation Questionnaire is the only one that incorporates almost all aspects of the mentioned theory (Silm et al., 2020).

Advantages of using self-reports to measure test-taking effort include: 1) they can capture a variety of non-effortful behaviour, 2) are easily applied in pencil-and-paper as well as computer-based test settings, 3) can be applied with any sample size (Ulitzsch et al. 2021).

The possible limitations of using self-reports are different response biases. As the questionnaire is usually positioned at the end of the test, test-takers may feel compelled to justify their results if they think that they did not do well or if they overestimate their motivation in order to fulfil the test-giver's expectation. Also, as the self-reports are mostly used as global measures (i.e. presented at the end of the test and asking about the whole test), it is difficult to evaluate one's motivation if it changed during the test-taking (Wise & Gao, 2017; Ulitzsch et al., 2021). Also, test-takers may not be fully aware of the motivational processes driving their behaviour.

2.4.2. Measures of TTM Based on Test-Taking Time

Test-taking time is directly observable and could be one measure that reflects effort and persistence. It can be reasoned that the test-takers who are willing to spend more time on a test (item) are more motivated. This interpretation is supported by the findings on speed-accuracy tradeoff, which show that the increase of test-taking speed comes at the expense of response accuracy (Heitz, 2014). This relationship, however, applies mostly in simple perceptual decision tasks and can be more complex with other type of tasks (Heitz, 2014; Ranger et al., 2021).

The presumed association between test-taking time and TTM and speed-accuracy tradeoff seem to contradict the finding that higher ability is related to shorter reaction times (Jensen, 2006). However, both can be true at the same time, because mental speed and test-taking speed are different constructs. Short response times not related to higher ability can also have several explanations. Birnbaum (1968) introduced the guessing parameter into IRT (item response theory) models, indicating that the test result is not only a function of item and individual parameters and that guessing or rapid responding can also have an impact. Rapid responding can be explained by a variety of reasons, such as low motivation (Wise & Kong, 2005) or time pressure (Schnipke & Scrams, 1997). Several measures and models based on response times have been proposed. Wise and Kong (2005) proposed the Response Time Effort (RTE) indicator (see below for more details). Later models have included IRT models for modelling test-taking effort (Liu et al., 2019; Ulitzsch et al., 2020; Ulitzsch et al., 2021).

RTE is a measure developed by Wise and Kong (2005) that indicates the amount of items that the test-taker answered effortfully. To calculate RTE, a time threshold is determined for each item; the threshold is the least amount of time that would indicate solution behaviour. If the answer is given in less time than the predetermined threshold, the response is viewed as rapid. To each item a dichotomous index of item solution behaviour is given: 1 if the response time exceeds the predetermined threshold, 0 if the answer time is less than the threshold. The RTE index characterizes the share of solution behaviour answers to all answers. RTE index of .9 means that for 90% of items solution-seeking behaviour was used, and 10% of the answers are considered rapid responses.

Studies that have used both RTE and SRE have found that there is a low to moderate correlation between the two measures (Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005), which means that RTE and SRE do not reflect the construct of test-taking effort in the same way. This raises the questions of what do either of these reflect more specifically and how can they be used in different testing contexts.

3. RESEARCH DESIGN AND METHODS

3.1. Educational Testing in Estonia

The empirical research described in this dissertation has been conducted in the Estonian context. Therefore, I will give a brief overview of tests Estonian students can come across. For the state to get feedback on how well the students are doing in their studies an external evaluation system was established in the mid-1990s (Tire, 2020). Main components of the system are sample based (standard-determining) tests for grades 3 and 6, and national examinations for grades 9 and 12 (Põhikooli- ja gümnaasiumiseadus, 2010). The standard-determining tests are centrally provided low-stakes tests and the students are not individually graded. The test results inform the schools and teachers what their students know and can do, and help to plan further teaching (HARNO, *n.d.*). All grade 6 tests are computer-based (Eurydice, 2020b; Eurydice, 2022).

In grade 9, at the end of compulsory education, students must take three centralized exams: in mathematics, Estonian language, and a subject chosen by the student (from a list of 10 subjects). The exams are assessed at the school by the subject teacher. The examinations are a necessary requirement for graduating basic school (Põhikooli- ja gümnaasiumiseadus, 2010, § 30).

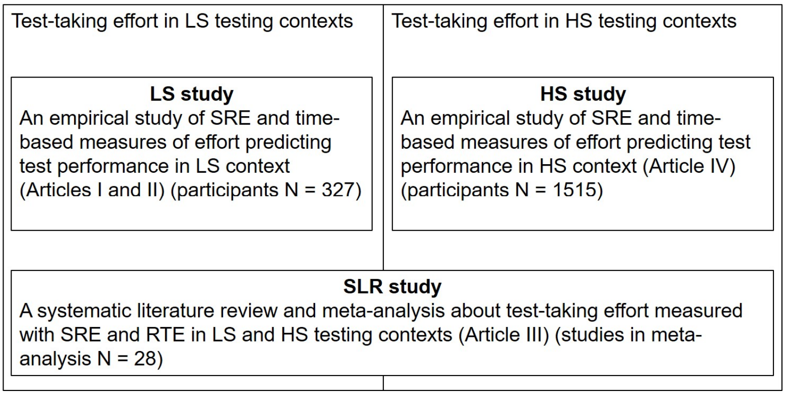
In grade 12, at the end of upper secondary school, students again must take three centrally set exams: in Estonian (or Estonian as a second language), mathematics, and a foreign language. The exams are centrally assessed (Põhikooli- ja gümnaasiumiseadus, 2010, § 31). The exam scores are often used by universities and other higher education institutions as a basis for admitting students. Therefore, the exams are usually high-stakes for the students. In addition to the centralised examinations, students need to pass a school exam and conduct an independent research project (Põhikooli- ja gümnaasiumiseadus, 2010, § 31).

Besides the centrally coordinated tests and exams, in everyday school life teachers test the students to evaluate their learning progress (Eurydice, 2022). Students can also come across admission tests. Some selective schools hold admission tests for accepting students to primary and/or to upper secondary school. Estonian universities and other higher education institutions stipulate different admission requirements; in most cases admission is decided on the basis of national examination results and/or an entrance exam or professional aptitude interview (Eurydice, 2020a). In the University of Tartu, the oldest and largest university in Estonia, also the Academic Test is employed (Must & Allik, 2002). Taking the Academic Test is compulsory for some specialties, and for most other specialties a good score on the test guarantees entrance to the university regardless of the scores received in the national examinations (Tartu Ülikool, 2022).

Since 2006 Estonia has also taken part in the OECD PISA study (Tire, 2020). It can be concluded that Estonian students can come across several LS (standard-determining tests, PISA study) and HS (graded tests, exams necessary for graduating, admission tests) testing situations.

3.2. Research Design

The research that forms the basis of this dissertation consists of three studies: 1) studying TTM in a LS testing context (LS study); 2) studying TTM in a HS testing context (HS study); 3) systematic literature review and meta-analysis (SLR study). The HS and LS studies are empirical studies in different testing contexts. In both studies we used time-based measures of effort and self-reported effort, and then created a model where the two effort indicators together predict test performance when other significant predictors of performance had been controlled for. In the SLR study we looked at the relationship between test performance and estimates of test-taking effort obtained with the two approaches (i.e. the time-based and self-reported indicators). The SLR study also enabled to give a state-of-the-art overview of the relationship between TTM (more specifically, test-taking effort) and performance. The last thorough synthesis of the research findings about the relationship between TTM and test performance was published by Wise and DeMars in 2005. Since then a considerable amount of new studies on TTM have been conducted. The three studies together provide a thorough understanding of how the estimates of test-taking effort obtained through two different approaches predict test performance. The design of the whole research is illustrated in Figure 2.



Test-taking effort in LS testing contexts

Test-taking effort in HS testing contexts

LS study
An empirical study of SRE and time-based measures of effort predicting test performance in LS context (Articles I and II) (participants N = 327)

HS study
An empirical study of SRE and time-based measures of effort predicting test performance in HS context (Article IV) (participants N = 1515)

SLR study
A systematic literature review and meta-analysis about test-taking effort measured with SRE and RTE in LS and HS testing contexts (Article III) (studies in meta-analysis N = 28)

Figure 2. *Overview of Research Design.*

3.3. Procedure and Participants

In both the LS and HS study data was obtained in authentic situations, i.e. the measures of TTM (including test-taking effort) were added to an ongoing study and an actual admission test, respectively. The LS study was a part of a study on first year university students' attitudes to learning, and academic ability. Their academic ability was assessed with a short version of academic test. The selfreport TTM questionnaire was added to the end of the test. As the test was computer-based, test-taking time on the item level was available and allowed computing the RTE index. In addition, the test-takers were asked to report their national examination results.

The HS study was carried out in the context of a university admission test (Academic Test). TTM questionnaire was added to the end of the Academic Test. As this test was also computer-based, it was possible to obtain time used per every item and calculate the RTE index. With the agreement from the Research Ethics Committee of the University of Tartu, we were able to retrieve the test-takers' national examination results from the institution responsible for the examinations. Sample description can be found in Table 1.

Table 1. *Sample Description for the LS and HS Studies.*

	LS study	HS study
Description	First year university students	University applicants
Sample size	327 (280)[a]	1515
Average age of the sample (years)	21.5 (SD = 2.1) (N = 280)	19.2 (SD = 2.8)
Gender (% of female participants)	73.7	61.6

Note. [a] Article I is based on the entire sample (N = 327), in Article II only data from the individuals who filled in the self-report questionnaire at the end of the test is used (N = 280).

A broader view of the measures estimating TTM was received via systematic literature review and meta-analysis. Within the systematic literature review we looked for articles in the context of education focussing on TTM or test-taking effort that also included some measure of TTM. In the meta-analysis we concentrated on zero-order correlations between test performance and test-taking effort. Three possible moderators (educational level, type of test-taking measure, and whether the test-takers were motivated) of the effect size were also added to the analysis.

3.4. Measures

In order to measure TTM we used a self-report questionnaire and test-taking time as indicators of test-taking effort. The self-report questionnaire was the Student Opinion Scale (SOS) (Sundre & Moore, 2002), which has been widely used for measuring TTM. The SOS is a post-survey questionnaire that consists of two subscales: the importance subscale (e.g., "Doing well on this test was important for me", "I am not curious about how I did on this test relative to others") and the effort subscale (e.g., "I gave my best effort on this test", "While taking this test, I was able to persist to completion of the task"). There are five statements in either subscale. The responses are on a five-point scale (1 – totally disagree, 5 – totally agree). In order to use the questionnaire in the Estonian context, the questionnaire was adapted into Estonian by the author of the dissertation.

As for test-taking time, this was recorded for each item separately. Different estimates based on test-taking time such as average time for correct answer, average time for incorrect answer, and RTE were calculated.

In the LS and HS studies the thresholds were obtained using an adaptation of the NT10 approach (Wise & Ma, 2012). According to the NT10 approach, 10% of the average time spent on all answers on an item is a suitable threshold for identifying rapid guesses. The adaptation made was that instead of all answers, only correct answers were involved. The rationale for this was based on the finding that rapid responding may be influenced by item positions, with more rapid responses near the end of the test (Weirich et al., 2017). Also, a minimum of 3 seconds and a maximum of 10 seconds were set for the thresholds.

TTM was measured when taking a short version of the academic test or the full-length Academic Test in the university admission setting. The Academic Test is similar to the Swedish Scholastic Aptitude Test (Wedman, 2017). It has seven subscales: vocabulary, diagrams, data sufficiency, text comprehension, mathematics, spatial ability, and foreign language comprehension. There are altogether 150 items to be solved within 180 minutes. Items are scored dichotomously. Missing answers are coded as incorrect answers. The score for the whole test is calculated based on subtest scores and standardised to a scale of 1–100, with a mean of 50 (SD = 16). The standardisation is based on the results of the same test used in a previous year.

The short version of the academic test was assembled by the author from Academic Test items from the years 2008– 2012. The short version consisted of only three subscales: vocabulary, mathematics, and spatial reasoning. The mean level of item difficulty was near 0.5 for all items. The test was conducted in an online research environment. Time limit for the test was 60 minutes, the average time used for taking the test was 37 minutes.

To make sure that the effect of TTM is not just a proxy of cognitive abilities (Penk & Richter, 2017; Wise & DeMars, 2005), results from previous HS testing situations, in the form of national examination results, were added in the models as control variables. In the LS study national examination results were self-reported. In the HS study national examination results were obtained from the Innove Foundation that carried out the national examinations in Estonia at the time. Most of the test-takers had scores for three exams, but there were some who had completed their examinations earlier when the requirements were different, and therefore had more or less examination results available. A mean score from the three highest exam results was calculated, to provide a general indicator of previous performance on HS tests.

3.5. Data Analysis

Within the empirical studies structural equation modelling (SEM) and bifactor analysis were used. SEM is an analysis method that enables to model the relationships between multiple observed and latent variables based on the structure of covariances between the variables. The observed variables represent the collected scores or other available data. The latent variables are hypothetical constructs or factors presumed to reflect a continuum not directly observable (Kline, 2015). In Articles II and IV SEM enabled to model performance and SRE as latent variables, predict performance with two different indicators of test-taking effort and control variables.

Bifactor analysis is considered to be an effective approach for "modelling construct relevant multidimensionality in a set of ordered categorical item responses" (Reise, 2012). A bifactor structural model specifies that the covariance among items can be accounted for by a single general factor reflecting a common variance in all scale items, and specific factor(s) that reflect additional common variance among clusters of items (Reise, 2012). Bifactor analysis was used in Article IV for determining whether the performance of the admission test can be viewed as unidimensional.

Within the systematic literature review, meta-analysis and moderator analysis were used in order to find out what kind of measures have been used to measure TTM, what is the average relationship between TTM and performance when using different measures, and what are the possible moderators of this relationship. A systematic review is "a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review" (Moher et al., 2009). For conducting the systematic literature review the instructions of PRISMA guidelines were followed (Moher et al., 2009).

Meta-analysis refers to synthesis of research that enables to "statistically combine the results of studies" (Cooper et al., 2009, p. 6). The original definition of meta-analysis coined by Glass (1976, as cited in Cooper et al., p.6) is "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings". The effect sizes from different studies combined with the meta-analysis often show more variability than expected based on only sampling error. A moderator analysis can explain whether any study descriptors are associated with the size of the effect (Cooper et al., 2009).

4. FINDINGS

4.1. In the LS Context SRE and RTE Complemented Each Other and Predicted a Significant Part of the Variance in the Test Results

In order to find how indicators of test-taking effort, both time-based and self-reported, predict performance in a LS testing context in relation to other important variables, SEM models were created. Test-taking effort was assessed with the adapted version of the Student Opinion Survey (Sundre & Moore, 2002), and based on test-taking time indicators such as time per item, average time for wrong answers, and RTE were calculated. Test-taking time appeared to be a good indicator of test-taking effort. It was evident that higher test-taking speed was related to lower performance (correlation between total test-taking time and test result, r = .716). It was also found that the average time for incorrect answers was shorter than the average time for correct answers ($M_{correct} = 52.5$, $M_{incorrect} = 48.8$, $t(279) = 3.2$, $p = .002$). This indicates that at least some wrong answers may be a result of rapid responding. One straightforward way to predict test performance with effort indicators is to use the number of answered items and time used per item as indicators of effort. We used this approach in Article I (see Figure 3). The model shows that time per item is almost as good predictor of test performance as the total number of items answered.

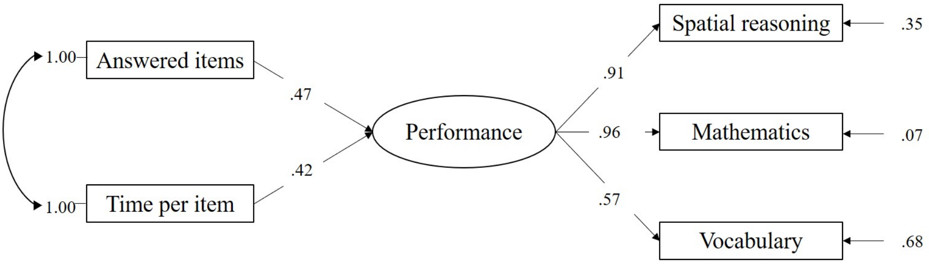


Figure 3. *Model Predicting Test Performance in a LS Testing Context with Number of Answered Items and Average Time per Item*s.

Note. $\chi 2$ (3) = 2.89, p = 0.41. Standardised solution. Model is from Article I.

In the further development of the model RTE and SRE were used as indicators of test-taking effort. The rationale behind this was that RTE gives more detailed information about effort on the test than total time for the test or average time per item, because for each item it is determined whether solution behaviour or rapid responding was used. Especially in online tests when we do not see what the test-taker is actually doing, total time per test may not be a good indicator. For example, a test-taker could possibly leave for 30 minutes and then return to the test, and the total test-taking time for them could be the same as for a test-taker who effortfully answered all of the items and was able to concentrate the whole time. The correlation between RTE and total test score was r = .71, and it was evident that when the test-takers with low motivation were filtered out from the dataset, the overall test results improved. For example, the mean score for the entire sample was 24.3 points, but when test-takers with RTE less than 0.9 were removed, the mean score became 28.2 (out of 45).

Another development in the model (Article II) involved incorporating an indicator of cognitive abilities (previous performance on a HS test), i.e. the average result from national examinations. Also, gender was added as a control variable. A SEM model with two indicators of test-taking effort (SRE and RTE) and control variables of previous performance in a HS testing situation (based on recalled national examination results) and gender predicted test performance. Test performance was modelled as a latent variable based on the three subtest scores of the short academic test. Also, SRE was modelled as a latent variable based on the answers to SOS effort subscale. Other variables were considered as observed variables.

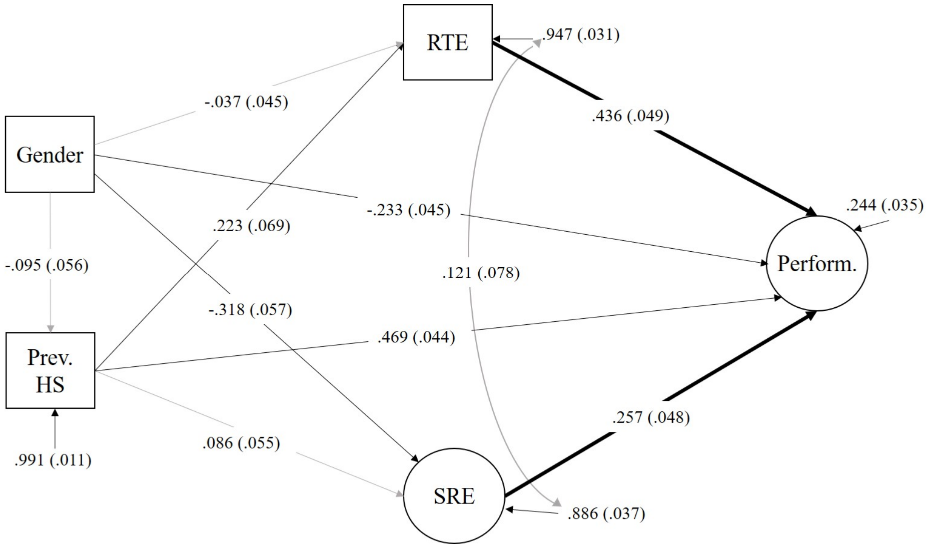


Figure 4. *Model Predicting Test Performance in a LS Testing Context with SRE and RTE.*

Note. Perform. – test performance; RTE – response time effort; SRE – self-reported effort; Prev. HS – previous high-stakes test result; Gender – 0-male, 1-female. Standardised solution. Thick lines represent the regression coefficients of motivational indicators on performance. Black lines represent statistically significant regression paths, grey lines represent statistically nonsignificant regression paths. $\chi2(37) = 102.58$, $p < .001$; RMSEA= .08; CFI = .95; SRMR = .05. Model is from Article II.

From the model it is evident that RTE is a stronger predictor of performance compared to SRE. But SRE also predicts test results separately from RTE. When only SRE is added to the model, the prediction of test performance improves by 8%; when only RTE is added, the prediction improves by 20%. Adding both effort indicators to the model approves the prediction of test performance by 25%. The model altogether predicts 75.6% of the variance in test performance. The correlation between SRE and RTE in the model is statistically nonsignificant.

It was also confirmed that RTE shows a ceiling effect (most values are near 1). Previous HS test performance predicted RTE, but not SRE. On the other hand, in the model gender predicted SRE, but not RTE. Both previous performance in a HS test and gender predicted test performance as well.

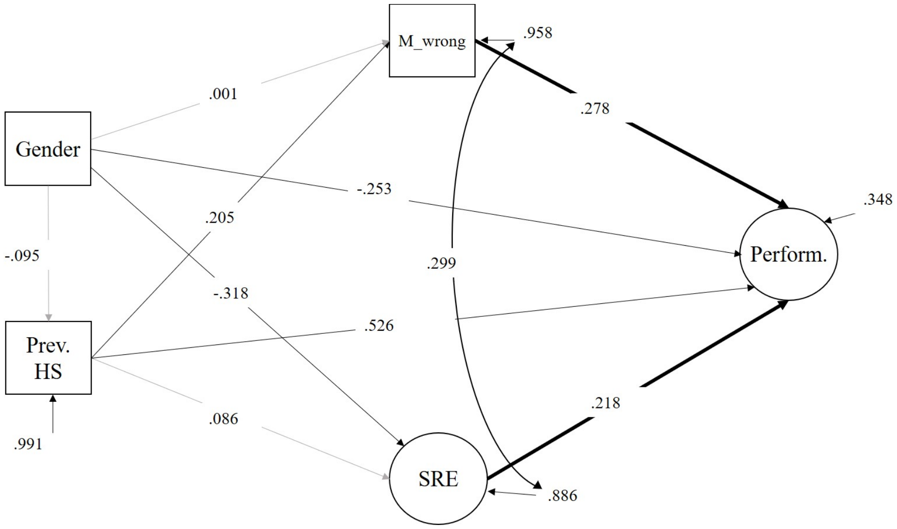


Figure 5. *Model Predicting Test Performance in a LS Testing Context with SRE and Average Time Spent on Incorrect Item.*

Note. Perform. – test performance; M_wrong – average time spent on incorrect item; SRE– self-reported effort; Prev. HS – previous high-stakes test result; Gender – 0-male, 1-female. Standardised solution. Thick lines represent the regression coefficients of motivational indicators on performance. Black lines represent statistically significant regression paths, grey lines represent statistically nonsignificant regression paths. $\chi 2(37) = 115.70$, $p < .001$; RMSEA= .09; CFI = .94; SRMR = .05. Model is from Article II.

Next to RTE, average time for wrong answer is also a good predictor of effort, but unlike RTE it does not show ceiling effect. It was supposed that wrong answers are given more quickly than right answers, presuming that right answers demand more time investment and wrong answers may be given due to hurrying. The model with average time for wrong answer predicted 65.2% of the variance in test performance (see Figure 5). The ceiling effect of RTE and strong correlation between RTE and test performance lead to the conclusion that RTE reflects better the bottom part of the motivation spectrum.

4.2. Test-Taking Effort had a Significant Relationship with Test Performance in the HS Context

A model similar to that used in the LS testing context was put to test in an authentic real-life HS testing context. Again, the SOS was added to the end of the test and test-taking time per item was logged for, which enabled calculating RTE and other time-based indicators. The models in LS and HS contexts are not directly comparable, but they provide information about the role of TTM in different testing contexts and the functioning of the different indicators of test-taking effort. If in the LS study the short version of the academic test consisted of three subscales with altogether 45 items, then in the HS study the Academic Test consisted of seven subscales with altogether 150 items. The Academic Test is intended to measure a unidimensional construct of academic ability. Bifactor analysis was used to see whether the assumption of unidimensionality is met. We concluded that the measured construct was essentially unidimensional, although one specific factor that had no clear meaning for us also emerged.

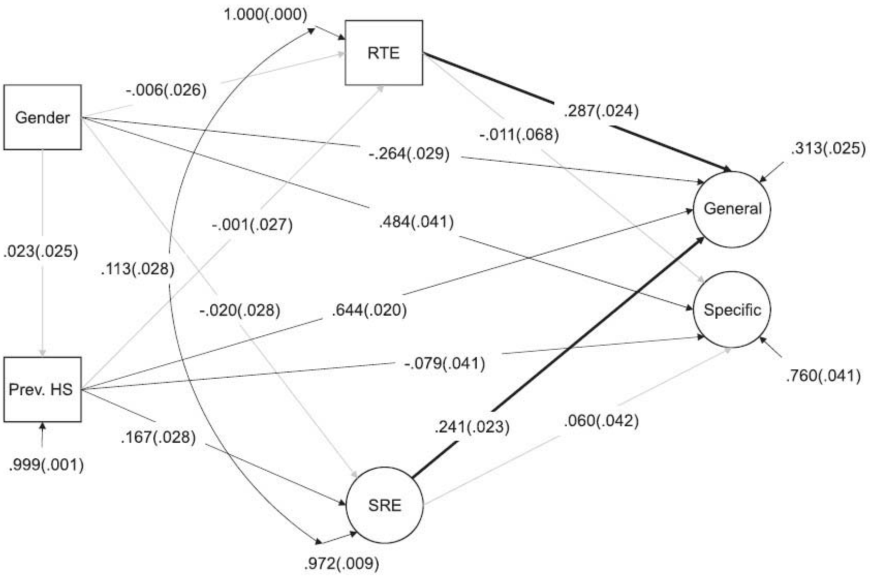


Figure 6. *Model Predicting Test Performance in a HS Testing Context with SRE and RTE.*

Note. General – general factor of performance; Specific – specific factor of performance; SRE – self-reported effort (SOS Effort subscale); Prev. HS – previous high-stakes test result (national examination results); Gender – 0-male, 1-female. Standardised solution. Thick lines represent the regression coefficients of motivational indicators on performance. Black lines represent statistically significant regression paths, grey lines represent statistically nonsignificant regression paths. $\chi2(73) = 401.97$, $p < .0001$; RMSEA= .055; CFI = .933; SRMR = .040. Model is from Article IV.

In the SEM model the general factor was considered as the indicator of performance. Similarly to the LS model, previous performance in a HS test (based on actual national examination results) and gender were added to the model. Performance and SRE were modelled as latent variables. It was found that the model predicted 68.7% of the variance in the general factor of test performance in the HS testing context. However, compared to the LS model previous HS test results predict larger part of the variance, and TTM indicators predict a smaller part of the variance in test performance compared to a LS testing situation, indicating that the variance of TTM in the LS context is larger than in the HS context, as can be expected (see Figure 6).

Again, both SRE and RTE predicted performance in the HS test. However, the effect of RTE was smaller than in the LS context. Average time for wrong answer did not correlate with performance in the HS context. Therefore, it seems that time-based indicators of TTM require more nuanced interpretation in HS testing contexts.

4.3. The Average Meta-Analytic Correlation Between Test-Taking Effort and Performance was Significantly Different When Measured Using SRE or RTE

As expected, the systematic literature review confirmed that TTM has been measured mainly in LS testing situations where it is viewed as a considerable threat to the validity of test score interpretation. Mostly self-report measures have been used, but time-based measures, such as RTE, have also been employed.

28 studies were identified that presented the zero-order correlation between test-taking effort and test performance. In several studies more than one effect size was presented, and therefore the meta-analysis was based on 55 effect sizes.

The average meta-analytic correlation between test-taking effort and test performance was .42. There was a large heterogeneity in the effect sizes (I^2 = 97.7%), and therefore moderator analyses were conducted. According to the literature and data available from the articles, three different moderators were tested for: 1) educational level, 2) type of TTM measure, and 3) motivating the test-takers. Educational level proved to be a significant moderator of the effect size between SRE and test performance, counting for about 35% of the heterogeneity. It was evident that for school students (grades 3–13) the relationship between SRE and test performance (r = 0.27, CI 0.22–0.30) is smaller than within university students (r = 0.39, CI 0.35–0.43). Whether the test-takers were somehow motivated before the test-taking did not have a moderating effect [Q(1) = 0.07, p = .79]. The same applies to the type of self-report measure, but the effects obtained with SRE were rather different from those obtained with RTE. The average meta-analytic correlation between performance and SRE was .33, and the correlation between performance and RTE was .72 (see Figure 7).

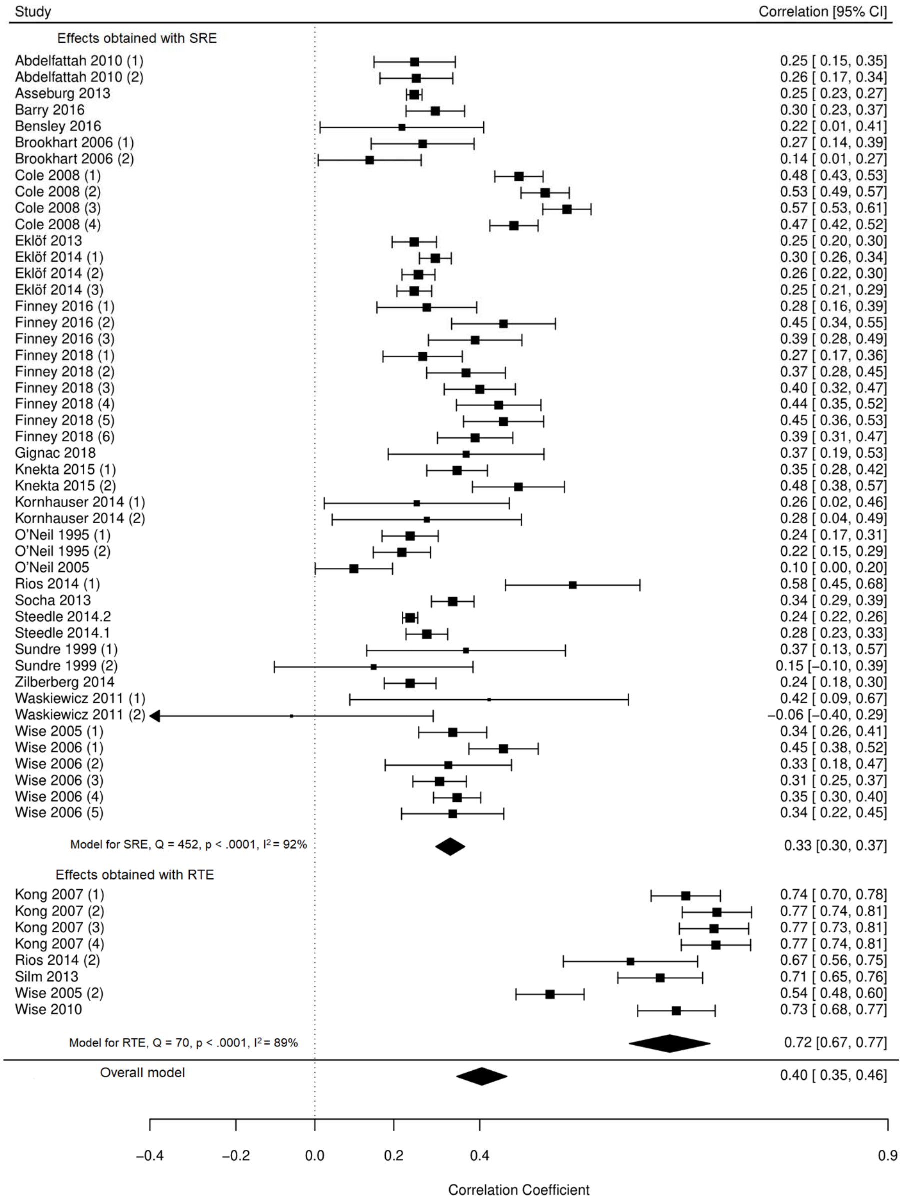


Figure 7. *Meta-Analytic Correlations Between Test-Taking Effort and Test Performance.*

Note. Squares mark the size of the correlation coefficient. Error bars mark the 95% confidence intervals. Diamonds represent the meta-analytic average correlation for studies using SRE and RTE respectively, and for all studies. References in the figure can be found in Article III.

5. DISCUSSION

Although the role of motivation in testing has been emphasised by early influential psychometricians (e.g. Cronbach, 1990; Spearman, 1927; Thurstone, 1919), the term "test-taking motivation" (TTM) has been coined in recent decades, and growing research interest has followed. TTM is considered important because it can lead to construct irrelevant variance (CIV) in test results and thus jeopardise the validity of the interpretation of the results. TTM has been measured primarily with several self-report instruments, but as these may be susceptible to various response biases, researchers Wise and Kong (2005) suggested the Response Time Effort index as an alternative. More precisely, RTE is a measure of test-taking effort. Various studies have shown that test-taking effort, compared to other components of TTM, is the best predictor of performance (Cole et al., 2008; Knekta & Eklöf, 2015; Penk & Schipolowski, 2015).

A few studies have shown that RTE and SRE distinguish between motivated and non-motivated test-takers in a relatively similar way (Swerdzewski et al., 2011). Most studies on the topic, however, show that the relationship between the two measures is relatively weak, indicating that they do not reflect the same thing (Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005). Nevertheless, both of the measures are able to predict performance (Silm, et al., 2020). The aim of this doctoral research was to investigate the two TTM indicators in more detail, to gain more information about TTM and its measurement. To this end, besides a literature review and meta-analysis, two empirical studies were conducted, one in a LS testing context and the other in a HS testing context. Based on the results three theses are proposed:

1) Time-based indicators are useful indicators of TTM, although they may be context-specific. As SRE may be biased by for example social desirability or self-justification, behavioural non-obtrusive indicators of TTM have been proposed. Wise and Kong (2005) created the RTE index which was also used in this thesis, although the use of time on test was studied in more detail. It was found that the average time for wrong answer could be another useful time-based indicator of test-taking effort. The rationale behind this is that a wrong answer can result from the perceived difficulty of the item or rapid responding. When the item is found difficult, but the test-taker takes time to solve it, it still reflects the effort put in, even if the attempt was unsuccessful. The research showed that time-based indicators of effort were better predictors of performance in a LS testing context compared to a HS testing context. There are several possible explanations. First, time-based indicators of effort show whether the test-takers refrain from rapid responding, but they do not distinguish between higher levels of motivation. Second, possible time pressure in the case of HS testing contexts can give RTE a different meaning – it can reflect either rapid responding or time pressure.

2) In addition to LS testing contexts, TTM manifests also in HS testing contexts. The studies in the dissertation showed that test-taking effort can be related to test results in both LS and HS testing contexts even when previous performance and gender are controlled for. Therefore, instead of strictly distinguishing between LS and HS testing situations and seeing a possible problem of low TTM only in LS contexts, it can be useful to consider the possible effect of TTM in all testing contexts. Nevertheless, the variance of TTM can be different according to specific testing context. Although the possible threat to test results validity due to low TTM is highest in LS testing contexts, the motivation depends on how individuals perceive the test and the context. For example, the HS testing context described in this dissertation was a university admission test. A university admission test can be considered a HS test, because it potentially has significant consequences for the test-taker. In the current case, however, the test was not the only way to gain admittance to the university, which explains the variance in TTM.

3) SRE and RTE characterise different parts of TTM and complement each other. It can be useful to assess TTM and test-taking effort with both self-report and time-based measures such as RTE, and interpret the results keeping in mind the context of the test. The meta-analysis included in the thesis showed that the average meta-analytic correlation was higher between test performance and RTE than between test performance and SRE. But the higher correlation does not necessarily show the superiority of RTE over SRE. The RTE index shows a ceiling effect with all values near 1. This means that RTE (at least when calculated based on the thresholds used in the current research) describes the lower end of the motivation spectrum and helps to distinguish the test-takers who clearly did not answer effortfully throughout the test. It provides no information about the upper part of the motivation spectrum: for example, there can be test-takers who take time to read the test items, but do not give their maximum effort when solving the items.

SRE has the potential for informing about the upper part of the motivation spectrum, but the possible biases of self-report data remain. In the current research, for example, it was evident that some test-takers reported low motivation, although their RTE index was high and they also received a high test result. When modelling RTE and SRE predicting performance, it was evident that both indicators had an independent effect on the test results in both the LS and HS testing context.

6. LIMITATIONS

The main limitation of the current thesis is that from all the aspects of TTM, it concentrates mostly on test-taking effort. Test-taking effort was chosen to be the focus of the current research proceeding from previous findings of effort being the best predictor of performance (Cole et al., 2008; Finney et al., 2018; Penk & Schipolowski, 2015; Zilberberg et al., 2014), and also because time-based measures like RTE capture mainly effort. Nevertheless, concentrating on only effort does not provide understanding of mental processes that precede effortful behaviour. For example, Finney et al. (2020) emphasise the role of emotions in the emergence of test-taking effort. Lundgren and Eklöf (2020) highlight that effort and motivation are not necessarily directly related; for example, there can be tasks that do not require much effort, but this does not necessarily mean that there is no motivation to do the task. Also, it has been shown that although in theory effort is not supposed to be related to the ability of the test-taker (Wise & Kong, 2005), in reality a positive relationship is often present (Wise, 2020). A limitation concerning the SLR study is that the ability of the test-takers was not considered, but zero-order correlations between test-taking effort and performance were used.

Another limitation concerns the use of the RTE indicator. Although RTE offers a valuable non-obtrusive behavioural alternative to self-reports, its use is dependent on the threshold determination method (Wise, 2019), which inevitably results in some amount of misclassifications due to differences in personal test-taking speed (Wise, 2020). Newest developments of modelling test-taking effort have attempted to also take into account the test-taker's cognitive ability. For example, the speed-accuracy+engagement model by Ulitzsch et al. (2020, 2021) uses Rasch modelling for that. However, even in this case the question remains as to which aspects of test-taking effort and motivation such models reflect. As for the usability of time-based measures, mostly researchers benefit from them, whereas for practitioners like school teachers they remain less practical. Not all assessments in schools are computer-based, and even if they were, analysing patterns of test-taking time is probably too time and labour intensive for the teachers.

Lastly, the empirical studies have been carried out in an Estonian context with university candidates or first year students as the sample, and the results may not be transferable to other contexts. It has also been shown that cultural differences may exist in TTM (e.g. Liu & Hau, 2020) and that TTM can be related to test-takers' age or school level that they attend (Rosenzweig et al., 2019; Silm et al., 2020).

7. CONCLUSIONS AND IMPLICATIONS

7.1. Theoretical Relevance

Findings presented in the dissertation indicate that TTM can be related to test performance in a HS testing context as well as in a LS context, although in the former the variance of TTM is smaller. This means that TTM should be recognized in all testing contexts, notwithstanding whether it is considered LS or HS, because the variance of TTM can differ depending on how the test-takers perceive the test. Drawing from the EV theory of motivation, it is logical to assume that motivation depends not only on the context but also on the internal cognitive and affective processes of the test-takers.

TTM can be assessed with different methods. Time-based measures of TTM have been proposed as alternatives to self-reported questionnaires, because the latter may not be objective. However, the analysis of the two methods showed that the methods cannot be used interchangeably. Rather, they describe different parts of the motivation spectrum and complement each other and can therefore give important information for more valid interpretation of the test results.

A meta-analysis of the performance difference in LS and HS tests was conducted by Wise and DeMars (2005) more than 15 years ago. The new literature review and the meta-analysis conducted in the context of the doctoral study gave a state-of-the-art overview of more recent findings on TTM; also, an average effect size of the relationship between test-taking effort and performance was presented.

7.2. Practical Relevance

Findings presented in the current dissertation support the notion that TTM should be considered and reported for a more valid interpretation of test results as proposed by the Standards (AERA, APA, & NCME, 2014). TTM can be considered in different phases of the research.

1) In the preparation phase the researchers/practitioners could acknowledge how the test-takers would possibly perceive the stakes of the test, i.e. whether the test results bear any consequence for them. In case there are no consequences, it is advisable to consider whether it is possible to raise the stakes of the test or stress the importance of it, so that the test-takers would be motivated to exert the effort necessary for their test score to reflect the construct that was intended to be measured.

2) When carrying out the study, it is possible to add a self-report measure of TTM to the test. If the test is computer-based it is also possible to log the test-taking times on item level, which in turn enables using test-taking time and the measures based on it (such as RTE) to be calculated and employed as indicators of TTM or, more specifically, test-taking effort. The test-taking time in general

provides a broad picture about test-taking effort, e.g. how much of the time allowed for test-taking was actually used or whether wrong answers were given more quickly than the correct ones. The latter may indicate that the wrong answers are a result of rapid responding. RTE can help to differentiate rapid responders from the rest of the test-takers. However, RTE does not differentiate the test-takers who did not practise rapid responding but were unmotivated in the sense that they could have done better if they put more effort in the test. Also, RTE can have different meanings in LS and HS testing situations. In HS testing situations it can also reflect time pressure – this can be concluded for example from where the rapid responses occurred and whether all of the allotted time was used. When all the rapid responses appeared in the end of the test and almost all of the time allotted for taking the test was used, then this indicates time pressure as the reason for lower RTE. Self-reported motivation and effort has the potential to give more information about the internal motivational processes, but has some significant drawbacks. Test-takers may not be aware of their internal motivational processes, they may not be motivated to answer the self-report questionnaire truthfully, or may answer in a socially desirable manner. Therefore, combining several indicators of motivation can give more information.

3) When interpreting test results, data about test-taking motivation should be considered. The Standards (AERA, APA, & NCME, 2014) suggest that alongside test results information about TTM should be reported. Also, the opportunity of filtering out data from unmotivated examinees can have an impact on the interpretation of test results.

Additionally, it is important to know that there can be age differences in the possible effect of TTM. For instance, the effect seems to be stronger in the case of university students compared to school students. Gender differences are also involved: women may be more willing to take a LS test, but they exert less effort and receive lower test scores.

REFERENCES

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. AERA.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. AERA.

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA seriously: How accurate are low-stakes exams? *Journal of Labor Research*, *42*(2), 184–243.

Apascaritei, P., Demel, S., & Radl, J. (2021). The difference between saying and doing: Comparing subjective and objective measures of effort among fifth graders. *American Behavioral Scientist*, *65*(11), 1457–1479.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*(4), 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*, 359.

Bandura, A. (1977). Self-efficacy: Towards a unifying theory of behavioural change. *Psychological Review, 84,* 191–215.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Boekaerts, M. (2002). The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. *Advances in Motivation and Achievement, 12,* 77–120.

Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). Motivation and test anxiety in test performance across three testing contexts: The CAEL, CET, and GEPT. *Tesol Quarterly, 48*(2), 300–330. https://doi.org/10.1002/tesq.105

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624.

Coluccia, E., & Louse, G. (2004). Gender differences in spatial orientation: A review. *Journal of Environmental Psychology, 24*(3), 329–340. https://doi.org/10.1016/j.jenvp.2004.08.006

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.

Cronbach, L. J. (1990). *Essentials of psychological testing.* Harper and Row publishers.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716–7720.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132.

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, 101859.

Eklöf, H. (2006). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In *The Second IEA International Research Conference* (Vol. 1, p. 135).

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569.

Eurydice (2020a, January 2)**.** *Bachelor's studies and studies in professional higher education. Estonia.* https://eacea.ec.europa.eu/national-policies/eurydice/content/bachelor-24_en

Eurydice (2020b, October 8). *Estonian standard-determining tests in Maths and Sciences take place online.* https://eacea.ec.europa.eu/national-policies/eurydice/content/estonian-standard-determining-tests-maths-and-sciences-take-place-online_en

Eurydice (2022, January 26). *Assessment in single-structure education. Estonia.* https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-single-structure-education-10_en

Finney, S. J., Myers, A. J., & Mathers, C. E. (2018). Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing, 18*(4), 297–322. https://doi.org/10.1080/15305058.2017.1396466.

Finney, S. J., Satkus, P., & Perkins, B. A. (2020). The effect of perceived test importance and examinee emotions on expended effort during a low-stakes test: A longitudinal panel model. *Educational Assessment*, *25*(2), 159–177.

Gagné, F., & St Père, F. (2001). When IQ is controlled, does motivation still predict achievement? *Intelligence*, *30*(1), 71–100.

Geisinger, K. F., & Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics. In Wells, C. S., & Faulkner-Bond, M. (Eds.), *Educational measurement: From foundations to future* (pp. 3–20). Guilford Publications.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, *8*(4), 369–395.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330.

Goldstein, H. (2014). Using league table rankings in public policy formation: Statistical issues. *Annual Review of Statistics and Its Application*, *1*, 385–399.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

HARNO (n.d.). *Tasemetööd.* Retrieved May 10, 2022. https://harno.ee/tasemetood

Heitz, R. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, 150. https://doi.org/10.3389/fnins.2014.00150

Holzinger, K. J. (1924). On scoring multiple response tests. *Journal of Educational Psychology*, *15*(7), 445.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53–69. https://doi.org/10.1037/0033-2909.104.1.53

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155. https://doi.org/10.1037/0033-2909.107.2.139

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier.

Johanson, M., Pedaste, M., Pastak, M., Täht, K., Sõrmus, M., & Jukk, H. (2021). Assessment of mathematical competence using the Estonian national e-tests. *Eesti Haridusteaduste Ajakiri. Estonian Journal of Education*, *9*(2), 100–126.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Knekta, E. (2017). Are all pupils equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with different stakes. *Scandinavian Journal of Educational Research, 61*(1), 95–111. https://doi.org/10.1080/00313831.2015.1119723

Knekta, E., & Eklöf, H. (2015). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire. *Journal of Psychoeducational Assessment, 33*(7), 662–673. https://doi.org/10.1177/0734282914551956

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., ... & Neubrand, M. (2005). Der mathematikunterricht der PISA-schülerinnen und -schüler. *Zeitschrift für Erziehungswissenschaft*, *8*(4), 502–520.

Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics, and science. *Educational Researcher, 18*(8), 17–27. https://doi.org/10.3102/0013189X018008017

Liu, Y., & Hau, K. T. (2020). Measuring motivation to take low-stakes large-scale test: New model based on analyses of "Participant-Own-Defined" missingness. *Educational and Psychological Measurement*, *80*(6), 1115–1144.

Liu, Y., Li, Z., Liu, H., & Luo, F. (2019). Modeling test-taking non-effort in MIRT models. *Frontiers in Psychology, 10*, 145. https://doi.org/10.3389/fpsyg.2019.00145

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation*, *26*(5-6), 275–301.

Messick, S. J. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). Macmillan Publishing Company.

Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis, 17*(4), 419–437.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.

Must, O., & Allik, J. (2002). *Tunne oma võimeid: abivahend eneseanalüüsiks*. [Know your abilities: a tool for self-analysis.] Tartu University Press.

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data?: A comparison of the reliability of data produced in graded and un-graded conditions. *Research in Higher Education*, *45*(8), 921–929.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: knowns and unknowns. *American Psychologist*, *51*(2), 77.

Núñez-Peña, M. I., Suárez-Pellicioni, M., & Bono, R. (2016). Gender differences in test anxiety and their impact on higher education students' academic achievement. *Procedia – Social and Behavioral Sciences, 228*, 154–160. https://doi.org/10.1016/j.sbspro.2016.07.023

OECD (2019a). How comparable are the PISA 2018 computer- and paper-based tests? In *PISA 2018 Results (Volume I): What Students Know and Can Do*, OECD Publishing. https://doi.org/10.1787/8f293551-en. [https://www.oecd-library.org/education/pisa-2018-results-volume-i_8f293551-en]

OECD (2019b). How much effort did students invest in the PISA test? In *PISA 2018 Results (Volume I): What Students Know and Can Do*, OECD Publishing. https://doi.org/10.1787/8f293551-en. [https://www.oecd-ilibrary.org/sites/04fd5153-en/index.html?itemId=/content/component/04fd5153-en]

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization, 115*, 94–110. https://doi.org/10.1016/j.jebo.2014.08.00.

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55–79.

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42,* 27–35. https://doi.org/10.1016/j.lindif.2015.08.002

Põhikooli- ja gümnaasiumiseadus (2010). https://www.riigiteataja.ee/akt/116042021007

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, *9*(1), 1–31.

Price, P. C. (2012). *Psychology research methods: Core skills and concepts.* Saylor Academy. https://saylordotorg.github.io/text_research-methods-in-psychology/index.html

Puksand, H. (2017). *Eesti keele riigieksam liigub e-eksami poole*. Emakeeleselts. https://www.emakeeleselts.ee/omakeel/2017_1/OK-1-2017_06.pdf

Ranger, J., Kuhn, J. T., & Pohl, S. (2021). Effects of motivation on the accuracy and speed of responding in tests: The speed-accuracy tradeoff revisited. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 15–38.

Reeve, C. L., & Lam, H. (2007). Consideration of g as a common antecedent for cognitive ability test performance, test motivation, and perceived fairness. *Intelligence*, *35*(4), 347–358.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen (Langversion, 2001). *Diagnostica*, *2*, 57–66.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 2014*(161), 69–82. https://doi.org/10.1002/ir.20068

Rosenzweig, E. Q., Wigfield, A., & Eccles, J. S. (2019). Expectancy-value theory and its relevance for student motivation and learning. In K. A. Renninger, & S. E. Hidi (Eds.). *The Cambridge handbook of motivation and learning* (pp. 617–644). Cambridge University Press.

Ryan, R. M. (Ed.). (2012). *The Oxford handbook of human motivation*. OUP USA.

Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology*, *85*(5), 739.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232.

Schunk, D. H., Meece, J. R., & Pintrich, P. R. (2014). *Motivation in education: Theory, research, and applications*. Pearson Higher Ed.

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335.

Spearman, C. (1927). *The abilities of man*. Macmillan.

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality, 22(*3), 185–209. https://doi.org/10.1002/per.676

Stenlund, T., Eklöf, H., & Lyrén, P.-E. (2017). Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice, 24(*1), 4–20.
https://doi.org/10.1080/0969594X.2016.1142935

Sundre, D. L. (1999, April 19–23). *Does examinee motivation moderate the relationship between test consequences and test performance?* Montreal, Quebec, Canada: Annual Meeting of the American Educational Research Association [Paper presentation].

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2

Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motivation. *Assessment Update, 14*(1), 8–9.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Tartu Ülikool. (2022, January 12). *Akadeemiline test.* https://ut.ee/en/node/115313

Thurstone, L. L. (1919). A method for scoring tests. *Psychological Bulletin, 16,* 235–240.

Tire, G. (2020). Educational Assessment in Estonia. In H. Harju-Luukkainen, N. McElvany, & J. Stang (Eds.), *Monitoring Student Achievement in the 21st Century* (pp. 119–129). Springer, Cham.

Tork, J. (1940). *Eesti laste intelligents: pedagoogiline, psühholoogiline ja sotsioloogiline uurimus.* Tartu Ülikool.

Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment*, *26*(2), 104–124.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*, 83–112.

Wedman, J. (2017). *Theory and validity evidence for a large-scale test for selection to higher education* [Unpublished doctoral dissertation]. Umeå universitet.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115–129. https://doi.org/10.1177/0146621616676791

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015.

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, *32*(4), 325–336.

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Conference presentation]. Annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65–83.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mentally taxing items. *Applied Measurement in Education, 8*(4), 341–351. https://doi.org/10.1207/s15324818ame0804_4

Yoakum, C. S., & Yerkes, R. M. (Eds.). (1920). *Army mental tests*. H. Holt. [https://archive.org/details/armymentaltests05yerkgoog/page/n15/mode/2up]

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, *13*(4), 519–552.

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing*, *14*(4), 360–384.

SUMMARY IN ESTONIAN

Testi täitmise motivatsioon madala ja kõrge olulisusega testimise kontekstides

Haridus- ja psühholoogiavaldkonnas kasutatakse teste mitmesuguste teadmiste, oskuste ja omaduste hindamiseks. Tihti ei pruugi aga testitulemused testitäitjate endi jaoks olulised olla. Sellest hoolimata võivad testitulemused omada tähtsust teistel tasanditel, näiteks kasutatakse neid erinevate õpilasgruppide võrdlemisel koolide või riikide lõikes. Kui aga saadav testitulemus ei ole testitäitja enda jaoks oluline ei pruugi tal testi täites olla motivatsiooni, et pingutada ja oma tegelikku võimekust või muud hinnatavat omadust demonstreerida. See on aga oluline tulemuste valiidsuse seisukohalt – kui testitäitjad ei pinguta, ei pruugi testi tulemused näidata ainult seda, mida sooviti hinnata, vaid tulemus peegeldab ka motivatsioonilist komponenti.

Kuigi tahte (sisuliselt motivatsiooni) olulisust testi tulemustes on rõhutanud kuulsad psühhomeetrikud juba ammu (Cronbach, 1990; Spearman, 1927; Thurstone, 1919), siis testi täitmise motivatsiooni (TTM) mõiste on teaduskirjandusse jõudnud viimastel aastakümnetel ja seda algselt personalivaliku kontekstis (Arvey et al., 1990). TTM-i uuringud on näidanud, et kõrge olulisusega testimise olukorras, kus testitäitja jaoks on testi täitmisel tagajärg (hinne, sissesaamine ülikooli vmt), on tulemused keskmiselt kõrgemad kui madala olulisusega testimise olukordades, kus selliseid tagajärgi ei ole. Wise'i ja DeMars'i (2005) meta-analüüsi järgi on vahe testi tulemustes keskmiselt 0,59 standardhälvet. Haridusliku ja psühholoogilise testimise standardid (AERA, APA, & NCME, 2014) soovitavad, et testi tulemusega koos esitataks teave ka TTM-i kohta. Üha enam seda ka tehakse, näiteks on TTM-i mõõdikud lisatud PISA uuringusse (Kunter et al., 2005; OECD, 2019b).

TTM-i nähakse peamiselt ootuse-väärtuse teooria raamistikus (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). Selle teooria kohaselt on sooritusele suunatud käitumine sõltuv edu ootusest ning subjektiivsest väärtusest, mida sooritus eeldatavasti kaasa toob. Mõlemad omakorda sõltuvad varasematest kogemustest ja omandatud hoiakutest. TTM-i on uuritud peamiselt madala olulisusega testimise olukordades, sest seal nähakse seda suurema ohuna tulemuste valiidsusele. Kuid lähtudes ootuse-väärtuse teooriast võib eeldada, et kuna motivatsioon ei olene ainult kontekstist, vaid ka näiteks inimese uskumustest, hoiakutest ja varasematest kogemustest, võib see mängida rolli ka näiliselt kõrge olulisusega testimise olukorras, kuna individuaalselt võidakse testi ja selle olulisust tajuda väga erinevalt. Vähesed uuringud, mis on TTM-i kohta kõrge olulisusega testimise olukorras läbi viidud, on andnud erinevaid tulemusi. Näiteks Sundre ja Kitsantas (2004) ei leidnud oma eksperimentaalses uuringus, et TTM oleks testitulemusega seotud. Küll aga on seose leidnud Knekta (2017), Knekta ja Eklöf (2015) ning Stenlund jt (2017). TTM-i mõju kõrge olulisusega testi tulemustes võib olla oluline näiteks juhul, kui tulemuste alusel toimub testiülesannete kalibreerimine.

Selleks, et TTM-iga arvestada ja tagada tulemuste valiidne tõlgendamine, on vaja TTM-i hinnata. Tavaliselt on seda tehtud enesekohaste küsimustikega (nt Student Opinion Scale, Sundre & Moore, 2002), kuid nendega saadud tulemused võivad olla mõjutatud mitmesugustest vastamiskalletest. Näiteks võib testitäitja vastata nii nagu tema arvates oodatakse, või vastata selliselt, mis õigustaks testis saadud tulemust (kui tulemus oli kehv, võib väita, et tegelikult ei pingutanudki). Kui testitäitja ei pinguta testi täites, on tõenäoline, et ta ei vasta küsimustikule läbimõeldult või ei soovi üldse vastata (Wise & Kong, 2005). Nende puuduste kõrvaldamiseks on Wise ja Kong (2005) pakkunud välja objektiivsema alternatiivi TTM-i hindamiseks, nimelt APV (ajaline pingutus vastamisel, ingl k. *Response Time Effort*) indeksi. Nimetatud indeksi arvutamisel võetakse arvesse, kui palju aega kulutas testitäitja etteantud testis iga ülesande lahendamisele ning võrreldakse tulemusi eelnevalt määratud lävedega. Lävi tähistab minimaalset aega, mis on vajalik, et etteantud ülesanne vähemalt läbi lugeda. Kui vastamiseks kulutatakse vähem aega, eeldatakse, et see vastus anti juhuslikult ja läbimõtlemata, mis viitab vähesele pingutusele ja madalale motivatsioonile. Ühest küljest on leitud, et APV-indeks ja enesekohaselt raporteeritud pingutus võimaldavad sarnaselt eemaldada andmestikust mittemotiveeritud testitäitjad (Swerdzewski et al., 2011). Teisalt on näidatud, et nende kahe tunnuse omavaheline seos jääb madalaks või mõõdukaks ning APV-indeksi puhul esineb laeefekt (Kong et al., 2007; Rios et al., 2014; Wise & Kong, 2005).

Käesoleva töö eesmärk oli uurida TTM-i võimalikku efekti nii madala kui ka kõrge olulisusega testimise olukordades, kasutades TTM-i hindamiseks nii enesekohaselt raporteeritud pingutust kui ka ajalisi pingutuse indikaatoreid (nagu APV-indeks) ning võrreldes nendega saadud tulemusi. Selleks püstitati järgmised hüpoteesid:

1) madala olulisusega teistimise kontekstis on TTM seotud testi tulemusega ka siis, kui kognitiivne võimekus ja sugu on arvesse võetud;
2) olukorras, mida võib pidada kõrge olulisusega testimise situatsiooniks, on TTM seotud testi tulemusega;
3) enesekohaselt raporteeritud pingutus ja ajapõhised pingutuse näitajad täiendavad üksteist ja suudavad kirjeldada iseseisvalt osa variatiivsusest testi tulemustes.

Hüpoteeside kontrollimiseks viidi läbi kolm uuringut:

1) empiiriline uuring madala olulisusega testimise olukorras;
2) empiiriline uuring kõrge olulisusega testimise olukorras;
3) kirjanduse ülevaade ja meta-analüüs testi täitmise pingutuse seosest testi tulemustega.

Leiti, et madala olulisusega testimise olukorras kirjeldab mudel, kus testi tulemuse ennustajaks on testi täitmise pingutus (nii enesekohane hinnang kui ka APV), varasem testitulemus (kognitiivse võimekuse indikaator) ja sugu, 75,6% testi tulemuse variatiivsusest. Sealjuures enesekohaselt hinnatud pingutus ja APV suurendasid ennustuse täpsust nii koosvõetuna kui ka eraldiseisvalt. Sarnane mudel ennustas tulemust ka kõrge olulisusega testimise olukorras, kuid pingutuse indikaatorid koos kontrollmuutujatega kirjeldasid võrreldes eelkirjeldatud mudeliga väiksema osa tulemuse variatiivsusest (68,7%). Meta-analüüsist ilmnes, et seose suurus pingutuse ja testi tulemuse vahel oleneb sellest, kas on kasutatud enesekohaselt hinnatud pingutust või APV-d. Seosed on vastavalt r = 0,33 ja r = 0,72.

Töö tulemuste alusel pakuti välja kolm teesi:

1) ajapõhised pingutuse näitajad võivad olla head TTM-i indikaatorid, kuid nende tõlgendus erinevates testimise situatsioonides võib olla erinev. Näiteks kõrge olulisusega testimise olukorras võib ajapõhine näitaja peegeldada mitte ainult vähese motivatsiooni, vaid ka ajasurve mõju;
2) kuigi TTM-i peetakse oluliseks peamiselt madala olulisusega testimise situatsioonides, võib see väljenduda ka näiliselt kõrge olulisusega testimise olukordades;
3) enesekohaselt hinnatud pingutus ja APV iseloomustavad motivatsioonispektri eri osi ning täiendavad teineteist.

Töö tulemused rõhutavad, et TTM-iga peab arvestama nii madala kui ka kõrge olulisusega testimise olukorras. Kuigi ajapõhiseid näitajaid on pakutud välja alternatiivina enesekohastele näitajatele, kirjeldavad nad vaid teatud osa motivatsioonispektrist. Pigem on tegemist teineteist täiendavate näitajatega, mis võivad mõlemad kaasa aidata testitulemuste valiidsemale tõlgendamisele.

ACKNOWLEDGEMENTS

My doctoral journey has taken place at one of the most exciting times in my life – in addition to the doctoral studies a real job at the University, starting a family, getting our first very own home. Also, during this time I have met many great and wise people. Here I would like to thank some of them and some of whom I knew before.

Indrek, with you I am happy. Thank you for the support and believing in me. Kirke and Fred – I am sorry the house has been such a mess. But, we will work on that more now that the dissertation is done. Gert and mom, thanks for helping out with everything. I guess you did not think I would be a student for that long.

My supervisors, there has been a lot to learn from you. Olev, thank you for always thinking along, guiding me to concentrate on meaning and content, and teaching me that there is always something to learn. I remember how discussing my preliminary ideas for a BA thesis lead to the idea of studying test-taking motivation, and then everything followed. Karin, thank you for the knowledge in data analysis. Also, thank you for responding to my letter more than 10 years ago and agreeing to supervise me with Olev. Margus, I admire your systematicity and energy. It is still a mystery for me how you do all the things you do. Thank you and Äli for welcoming me in the Institute of Education. Äli, although not my doctoral supervisor, you have always been very supportive and sharing your good ideas.

Liina A., Pihel, Liina M., Triin – you have been the best part of doctoral studies. All different, yet good (and fun) together. Liina – I admire your confidence and diligence, Pihel – your social skills and sense of humour, Liina – your organizational skills and courage, Triin – your mastery in self-regulation and positive outlook on life. Thank you all for the friendship!

My friends from before the doctoral studies, I am sorry I have been distant. Iris and Triin, thanks for sticking with me. Margot, thank you for reading the dissertation with a critical eye – I wish I was so observant as you are.

PUBLICATIONS

CURRICULUM VITAE

Name:	Gerli Silm
Date of birth:	17.12.1986
Citizenship:	Estonian
Work address:	University of Tartu, Faculty of Social Sciences, Institute of Education, Jakobi 5, Tartu 51005
Phone:	+372 5647 9324
E-mail:	gerli.silm@ut.ee

Education:

2015–….	University of Tartu PhD studies in Educational Science
2012–2015	University of Tartu Master studies in Psychology (*cum laude*)
2009–2012	University of Tartu Bachelor studies in Psychology
2006–2009	University of Tartu Pärnu College Diploma in Tourism and Hotel Management
1997–2006	Tallinn Väike-Õismäe Secondary School (*graduated with gold medal*)
1994–1997	Private school Maikool

Professional employment:

2020–…	University of Tartu, Institute of Education Specialist of Educational Science
2015–2020	University of Tartu, Institute of Education Junior Researcher

Publications:

Articles

Silm, G., Must, O., Täht, K., & Pedaste, M. (2021). Does test-taking motivation predict test results in a high-stakes testing context? Educational Research and Evaluation, 1–27.

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. Educational Research Review, 31, 100335.

Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. TRAMES: A Journal of the Humanities & Social Sciences, 23(3).

Silm, G., Tiitsaar, K., Pedaste, M., Zacharia, Z. C., & Papaevripidou, M. (2017). Teachers' Readiness to Use Inquiry-Based Learning: An Investigation of Teachers' Sense of Efficacy and Attitudes toward Inquiry-Based Learning. Science Education International, 28(4), 315–325.

Hunt, P., Leijen, Ä., Silm, G., Malva, L., & Van der Schaaf, M. (2016, October). Student Teachers' Perceptions About an E-portfolio Enriched with Learning Analytics. In International Computer Assisted Assessment Conference (pp. 39–46). Springer, Cham.

Pedaste, M., Must, O., Silm, G., Täht, K., Kori, K., Leijen, Ä., & Mägi, M. L. (2015). How do cognitive ability and study motivation predict the academic performance of IT students?. In ICERI conference.

Silm, G., Must, O., & Täht, K. (2013). TEST-TAKING EFFORT AS A PREDICTOR OF PERFORMANCE IN LOW-STAKES TESTS. TRAMES: A Journal of the Humanities & Social Sciences, 17(4).

Reports

Silm, G., Tiitsaar, K., Valk, A. (2021). Kõrghariduse rahastusmudelid ja nende muutmise võimalikud mõjud. Teaduskirjanduse analüüs. Tartu Ülikool.

Silm, G. (2019). Eesti õpilaste heaolu ja õppimisega seotud hoiakud. PISA 2018 Eesti tulemused (lk 130–150). Tallinn.

Täht, K., Silm, G. (2017). Eesti õpilaste meeskondlik ehk koostöine probleemilahendusoskus PISA 2015 näitel: Tartu

Valk, A., Silm, G. (2015). Haridus ja oskused: PIAAC uuringu temaatiline aruanne nr 6. Tartu: Haridus- ja Teadusministeerium.

Chapters

Silm, G., Täht, K., Must, O. (2013). Testi täitmise motivatsiooni mõju testi tulemustele. Kõrgkool ja psühholoogia (lk 78−94). Tartu Ülikool.

Silm, G. (2016). Lapsevanemate tagasiside Tartu koolides. Rõõmuga kooli: kodu ja kooli koostöö käsiraamat (lk 91–104). Tartu.

ELULOOKIRJELDUS

Nimi:	Gerli Silm
Sünniaeg:	17.12.1986
Kodakondsus:	Eesti
Töökoha aadress:	Tartu Ülikool, sotsiaalteaduste valdkond, haridusteaduste instituut, Jakobi 5, Tartu 51005
Telefon:	+372 5647 9324
E-post:	gerli.silm@ut.ee

Haridustee:

2015–….	Tartu ülikool haridusteaduste doktoriõpe
2012–2015	Tartu ülikool Psühholoogia magistrikraad (*cum laude*)
2009–2012	Tartu ülikool Psühholoogia bakalaureuse kraad
2006–2009	Tartu ülikooli Pärnu kolledž Turismi- ja hotelliettevõtluse diplom
1997–2006	Tallinna Väike-Õismäe gümnaasium (*lõpetatud kuldmedaliga)*
1994–1997	Erakool Maikool

Töökogemus:

2020–…	Tartu ülikool, haridusteaduste instituut haridusteaduste spetsialist
2015–2020	Tartu ülikool, haridusteaduste instituut nooremteadur

Publikatsioonid:

Teadusartiklid

Silm, G., Must, O., Täht, K., & Pedaste, M. (2021). Does test-taking motivation predict test results in a high-stakes testing context? Educational Research and Evaluation, 1–27.

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. Educational Research Review, 31, 100335.

Silm, G., Must, O., & Täht, K. (2019). Predicting performance in a low-stakes test using self-reported and time-based measures of effort. TRAMES: A Journal of the Humanities & Social Sciences, 23(3).

Silm, G., Tiitsaar, K., Pedaste, M., Zacharia, Z. C., & Papaevripidou, M. (2017). Teachers' Readiness to Use Inquiry-Based Learning: An Investigation of

Teachers' Sense of Efficacy and Attitudes toward Inquiry-Based Learning. Science Education International, 28(4), 315–325.

Hunt, P., Leijen, Ä., Silm, G., Malva, L., & Van der Schaaf, M. (2016, October). Student Teachers' Perceptions About an E-portfolio Enriched with Learning Analytics. In International Computer Assisted Assessment Conference (pp. 39–46). Springer, Cham.

Pedaste, M., Must, O., Silm, G., Täht, K., Kori, K., Leijen, Ä., & Mägi, M. L. (2015). How do cognitive ability and study motivation predict the academic performance of IT students?. In ICERI conference.

Silm, G., Must, O., & Täht, K. (2013). TEST-TAKING EFFORT AS A PREDICTOR OF PERFORMANCE IN LOW-STAKES TESTS. TRAMES: A Journal of the Humanities & Social Sciences, 17(4).

Aruanded

Silm, G., Tiitsaar, K., Valk, A. (2021). Kõrghariduse rahastusmudelid ja nende muutmise võimalikud mõjud. Teaduskirjanduse analüüs. Tartu Ülikool.

Silm, G. (2019). Eesti õpilaste heaolu ja õppimisega seotud hoiakud. PISA 2018 Eesti tulemused (lk 130–150). Tallinn.

Täht, K., Silm, G. (2017). Eesti õpilaste meeskondlik ehk koostöine probleemilahendusoskus PISA 2015 näitel: Tartu

Valk, A., Silm, G. (2015). Haridus ja oskused: PIAAC uuringu temaatiline aruanne nr 6. Tartu: Haridus- ja Teadusministeerium.

Peatükid kogumikes

Silm, G., Täht, K., Must, O. (2013). Testi täitmise motivatsiooni mõju testi tulemustele. Kõrgkool ja psühholoogia (lk 78−94). Tartu Ülikool.

Silm, G. (2016). Lapsevanemate tagasiside Tartu koolides. Rõõmuga kooli: kodu ja kooli koostöö käsiraamat (lk 91–104). Tartu.

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.

37. **Tõnis Männiste**. Measuring military commanders’ decision making skills in a simulated battle leading environment. Tartu, 2020, 256 p.
38. **Maile Käsper.** Supporting primary school students’ text comprehension and reading interest through teaching strategies. Tartu, 2021, 158 p.
39. **Liina Malva**. Teachers’ General Pedagogical Knowledge: Its Nature, Assessment and Representation in Practice. Tartu, 2021, 163 p.
40. **Liina Adov**. Predicting teachers’ and students’ reported mobile device use in stem education: The role of behavioural intention and attitudes. Tartu, 2022, 160 p.
41. **Wilson Ofotsu Otchie.** Social Media in Education: Contextualizing Teaching with Social Media in High School. Tartu, 2022, 205 p.
42. **Karmen Kalk.** Using blogs to promote and predict reflection during teaching practice and induction year. Tartu, 2022, 139 p.