TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Kristian Kivimägi 143907IVCM

# PREDICTING STUDENTS' SUCCESS USING TECHNICAL LABS AS PART OF UNIVERSITY ADMISSION TO A CYBER SECURITY PROGRAM

Master's Thesis

Supervisors: Kaie Maennel
Olaf Manuel Maennel

Tallinn 2019

# Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Kristian Kivimägi

11.12.2019

# Abstract

Due to the existing and predicted future skills gap in the cybersecurity labour market, the educational institutions are establishing and admitting the future specialists to the cybersecurity study curricula. The admission boards have to select from the large numbers of applicants with different backgrounds who are interested in studying cybersecurity and to minimize later dropouts and study times. Thus, valid admission and selection procedures are critical, however, there is a lack of scalable (i.e., minimising efforts of human evaluator) and validated admission procedures that evaluate cybersecurity technical skills.

This thesis aims to validate innovative and scalable university admissions process that includes technical skills assessment using the cloud-based and remote virtual labs to assess whether the labs used can be used as predictor for the future success in the study program. The research uses data collected during admission procedure and data collected from technical courses during the first semester. As limited demographic or historic study behavioural data is available about the students, the developed prediction model is limited to use only information from the admission process and assessments at the start of studies.

The results suggest that technical skill assessment can be used as a significant predictor to assess potential candidates' skill level. However, the interview score is not redundant either. Instead, both of the admission methods are rather complementary to each other, both methods address different, but equally relevant aspects of student performance.

The main contributions of this thesis is the validation whether online technical labs on the individual cybersecurity topics can be used as significant predictor to measure potential candidates' skill level and later success in technical courses.

This thesis is written in English and is 48 pages long, including 6 chapters, 7 figures and 14 tables.

# Annotatsioon

# Õpilaste edu prognoosimine ülikooli vastuvõtu tehniliste testide põhjal küberkaitse õppesuunal

Tulenevalt juba olemasolevast ja prognoositavast oskuste puudujääkidest küberturbe tööturul, akadeemilised institutsioonid koolitavad uusi küberturbe spetsialiste. Vastuvõtu komisjon peab valima suure hulga ja erineva taustadega kandidaatide seast inimesed, kes on huvitatud küberturvet õppima samal ajal minimeerides hilisemaid väljalangejaid õppeprogrammist. Seega on kriitiline, et ülikooli vastuvõtu protseduurid oleksid täpsed, kuid hästi skaleeruvaid ning tehnilisi oskusi hindavaid vastuvõtu protseduure praktiliselt ei eksisteeri.

Selle magistritöö eesmärk on valideerida uuendusliku ja skaleeritavat ülikooli vastuvõtuprotsessi, mis hõlmab tehniliste oskuste hindamist pilvepõhiste virtuaalsete laborite abil. Uurimistöös kasutatakse vastuvõtuprotseduuri ajal kogutud andmeid ja esimesel poolaastal tehnilistelt kursustelt kogutud andmeid. Kuna õpilaste kohta on piiratud kogus demograafilisi ja ajaloolisi käitumisandmeid, siis mudel peaks kasutama ainult sisseastumisprotsessist ja õpingute alustamisel saadud teavet.

Lõputöö tulemused viitavad sellele, et tehnilisi teste saab kasutada olulise ennustajana potentsiaalsete kandidaatide oskuste taseme hindamisel. Intervjuu tulemus on samuti vajalik. Seega, mõlemad vastuvõtu protsessis kasutatavad meetodid on üksteist täiendavad. Mõlemad meetodid hindavaid erinevaid, aga asjakohaseid aspekte õpilase õpitulemustes.

Selle lõputöö peamine panus on valideerimine, kas küberturbe tehnilisi laboreid saab kasutada olulise enneustajana potentsiaalsete kandidaatide oskuste taseme ja hilisema õppeedukuse hindamisel.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 48 leheküljel, 6 peatükki, 7 joonist, 14 tabelit.

# List of abbreviations and terms

| | |
|---|---|
| MSc | Master of Science |
| CST 1 | Cyber Security Technologies 1 |
| CST 2 | Cyber Security Technologies 2 |
| WASE | Web Application Security |
| LA | Learning Analytics |
| ECTS | European Credit Transfer and Accumulation System |
| CV | curriculum vitae |
| HTML 5 | Hypertext Markup Language 5 |
| Mbps | Megabit |
| HTTPS | HyperText Transfer Protocol Secure |
| SQL | Structured Query Language |
| SOC | Security Operations Centre |
| STEM | Science, technology, engineering and mathematics |
| RQ | Research question |

# Table of contents

# List of figures

# List of tables

# 1 Introduction

## 1.1 Thesis Motivation

The number of internet-connected services, computers and servers are steadily increasing with each day. The number of internet-connected devices reached 22 billion at the end of 2018 [1] and is estimated to reach 38.6 billion devices by 2025 [2]. This means that the attack surface that hackers can potentially abuse is increasing each year. In addition to the increasing attack surface, cybercriminals are becoming more innovative and cyberattacks are evolving and becoming more technically sophisticated, this, in turn, makes them more impactful and also avoid detection [3].

In order to combat this threat well educated cyber security professionals are needed. However, the existing workforce is not enough to fight the criminals and the demand for new specialists is high [4]. To address this issue the academic institutions have established cybersecurity programs to train more specialists [5]. This is important as cybersecurity professionals are a vital component in combating cyber threats [6] and they are required to have a high level of competency to create and implement technologies, as well as manage human resources in order to: identify cyber threats and vulnerabilities, protect information and resources, detect the occurrences of cybersecurity events, respond to incidents, as well as recover from cybersecurity events [7].

This, in turn, introduces a new challenge - how to ensure that admission or human resources hiring processes selects the candidates who will be likely to succeed in their later studies or are suitable specialists for the tasks? In order to select the best and brightest admission procedure consists of assessments to determine the student's technical skill level. Traditionally the assessments are knowledge-based however these assessments may exclude suitable candidates [8]. Traditional assessments measure what a person knows about a subject matter, but this method does not unveil what a person can actually do. Skills and knowledge need to be assessed in a balanced way - therefore knowledge questions and tasks both must have their place in an assessment procedure. This is one of the reasons why hands-on technical labs are gaining popularity. To no surprise, they are

also gaining popularity in the universities cyber security programs. Such labs are used for both in assessing students' skills and in the teaching process. One of the main motivations for introducing technical labs in the admission process is the reason to allow a candidate who may not be strong in the interview context, show off their technical skills instead. For example, a student who is very nervous during an interview but showed skills in labs can be detected [8].

TalTech University has an international Cyber Security master's program that is being jointly operated with the University of Tartu. Admittance to the study program is based on admission threshold. The candidate must provide a motivational letter, a bachelor's degree, other relevant documents and a fee for admission. In addition, the admission process consists of an online interview and online technical assessment [9]. The technical assessment presented to applicants is based on Rangeforce platform. The company provides scalable platform to train people in the cyber security field [10].

Currently, the technical assessments are an optional part of the admission process. If the candidate chooses to complete the online technical assessment the admission board can skip the technical questions during the interview, as the candidate has already demonstrated the skill during the online technical assessment. The admission board is interested to validate whether the online technical assessment used is an accurate predictor to rank potential candidates. If the assessment proves to be a valid predictor the admission board might consider making the labs as a mandatory step during the admission process.

## 1.2 Problem Statement

The admission boards have to select from large numbers of applicants with different backgrounds who are interested in studying cybersecurity and to minimize later dropouts and study times. Thus, valid admission and selection procedures are critical, however, there is a lack of scalable and validated admission procedures that evaluate cybersecurity technical skills. For this purpose, the online technical assessments, could be used. However, it must be verified that the online technical assessments are a valid predictors of student performance and measure the technical skill level appropriately.

The thesis author is the instructor for Cyber Security Technologies, a mandatory course for 1st-year MSc students. To understand what course a student should take, there is an online technical assessment (Web Application Security assessment) that is conducted during the first lecture. This test determines students' current skill level and which of the two courses the student can take.

To validate whether these online technical assessments are valid predictors of student performance more analysis is needed to understand the skill of each individual student.

The research collects data during admission procedure and from technical courses during the first semester to validate whether online technical assessments could be used to predict student future performance. The constraints applied in the model are that it should use only information from the admission process and assessments at the start of studies, as limited demographic or historic study behavioural data is available about the students.

### 1.2.1 Research Questions

The main research question that this study addresses is:

Can online hands-on technical assessment labs be used as a significant predictor to assess potential candidates' skill level and future success?

The specific research questions this research study addresses are:

1. Can online technical assessment labs be used as a significant component in predicting a candidate's future performance? In particular:

    1.1. How do interview and hands-on technical assessment labs components of the admission and selection process and their results predict student performance?

    1.2. Does the Web Application Security (WASE) assessment and its results predict student performance?

    1.3. Is the WASE assessment necessary, if the online technical assessment used in the admission process predict student performance?

2. Is the course of an appropriate difficulty that is neither too easy nor too difficult?

### 1.2.2 Research Approach

The author is analysing the applicants' admission process and course completion data and applying it for prediction for student success. Predictive analytics is used in combination with historical data and statistical algorithms to identify the likelihood of future outcome. While descriptive statistics help researchers understand what happened, predictive models aim to answer why it happened [11]. Linear regression is used to understand key relationships and why it happened. Later this can be used to predict values and apply the same algorithm for next year students. In overall research approach, the author follows learning analytics model [16].

### 1.2.3 Contribution

The main contribution of this paper is the validation whether Rangeforce[1] or similar online technical labs on the individual cybersecurity topics can be used as significant predictor to measure potential candidates' skill level and later success in technical courses.

The thesis will provide a model to predict student performance. The model is described in the section "Research method" and it provides guidance for other researchers who can use this method of predicting student performance with their own data.

While there are existing methods that have been used in the admission procedure to predict the student future study success, there are no existing methods that are fit to be used with Rangeforce online technical assessments.

In addition, the thesis author designed cyber security technologies (CST) course and is one of the instructors for the course. The research paper gives feedback to the course instructors whether the method used to assign students to CST I and CST II is appropriate and if the home assignments are of appropriate difficulty for students.

[1] https://rangeforce.com

# 2 Thesis Background

This chapter gives background information that is relevant to understand the terms and context that will be used throughout the thesis. The author provides overview of TalTech's Cyber Security Master's program, Cyber Security Technologies course and learning analytics.

## 2.1 Program Overview

TalTech Cybersecurity Masters curriculum nominal duration is 2 years and the study load is 120 ECTS. The curriculum consists of the following modules: general studies, core studies, special studies, free-choice courses and graduation thesis. During the studies, the students have to choose a speciality. There are three specialities in the Cybersecurity Masters Programme: Cybersecurity, Digital Forensics and Cryptography. All specialities have to complete general studies, core studies, Cybersecurity Technology, selection of free choice courses, speciality courses and write a master's thesis to graduate from the programme [Table 1]. Graduates will receive a joint Diploma signed by both TalTech and Tartu University [12].

Table 1. Model Overview

| **Modules** | **Speciality** | **ECTS** |
|---|---|---|
| General studies | Cyber Security, Digital Forensics, Cryptography | 6 |
| Core studies | Cyber Security, Digital Forensics, Cryptography | 24 |
| Cyber Security Technology | Cyber Security, Digital Forensics, Cryptography | 6 |
| Special Studies in Cyber Security | Cyber Security | 54 |
| Special Studies in Digital Forensics | Digital Forensics | 54 |

| | | |
|---|---|---|
| Special Studies in Cryptography | Cryptography | 54 |
| Free Choice Courses | Cyber Security, Digital Forensics, Cryptography | 6 |
| Graduation Thesis | Cyber Security, Digital Forensics, Cryptography | 24 |

## 2.2 Admission Process

Each applicant for the cybersecurity master's program has to go through the following steps: [8]

- Submit their CV, a motivation letter and other relevant documents through the online application platform
- Fill in self-evaluation questionnaire
- Complete at least one out of four different virtual labs. The labs are available in a two-week period.
- Online interview via Skype that lasts around 10-15 minutes

### 2.2.1 Technical Assessment

Virtual labs presented to applicants are based on Rangeforce platform.  platform enables on-demand access to virtual lab environment via a web browser. [13] This setup means that an applicant only needs an HTML5 capable web browser do access and a decent Internet connection (at least 3Mbps). Rangeforce system provides a Virtual Teaching Assistant which guides and provides participants hints. It is impossible to measure all the skills of the participants, so the exercises are chosen to represent a mix of technical topics. [14] The selected labs are the following: [8]

- Introduction lab—essential command line skills (Git, apt-get, Apache server). Estimated completion time 25 minutes;
- HTTPS Security—basic level skills connected to command line, public key infrastructure, and server administration basics; estimated completion time 45 minutes;

- SQL injection—intermediate level skills connected to attacking SQL databases (SQL, SQL injection); estimated completion time 90 minutes;
- Botnet—advanced level skills connected to network scanning skills, text parsing (programming skills are beneficial) and SQL injection skills; estimated completion time 45 minutes.

The choice of these different exercises is based on typical attack vectors that students are likely to encounter in their future cybersecurity jobs and require different skill levels (from essential to advanced). The combination of exercises is used to determine the level of skill, but also to cover a variety of different skills. Each lab has a predetermined skill level from basic to advanced level [8].

### 2.2.2 Admission Interview

Interviews conducted during the admission process are usually 10 – 15 minutes long. The interview starts with an introduction where the candidate speaks why about himself/herself and why the individual chose to apply to TalTech. The interview then transmissions to technical questions. The aim is to see the candidate's knowledge and logical thinking. The interview ends with a closing note and explanation of next steps [8].

## 2.3 Cyber Security Technologies Course

Information Security has emerged from being a specialist topic that was studied mainly by military and governmental agencies to a general subject that is relevant to all professionals who are part of developing or using modern-day information and communication systems. Most courses taught in the information security field emphasize theory and touch basic concepts of cryptography, algorithms, protocols and models [15]. However, information security is ultimately about getting your hands dirty and putting these ideas to use.

The goal of Cyber Security Technologies course is to provide a hands-on experience and achieve a coherent understanding of technology for the students. They need to have a basic understanding of cybersecurity technology. The lectures consist of a theory part and laboratory assignments about the same subject they listen in the lecture. That way they are directly involved and see the practical consequence [15].

The main focus of the courses is related to what tools and methods can be used to secure networks, operating systems and web applications. Each of these topics are wide and could be several different courses on their own, but the goal of the subject is to introduce the 1st year students to these topics so that they understand them and can have an intelligent discussion about them. The technologies introduced in the courses are mostly open source with a couple of exceptions. Open-source security tools often require more time for initial setup and administration, but they are a great place to get started and set up initial security controls for the organisation.

All the students in the cyber security master's programme must attend either Cyber Security Technologies 1 or Cyber Security Technologies 2 course. Both courses are designed for 1st-year students. The experience and background of enrolled students vary quite a lot, there are students who do not have an IT background, some who have limited experience and people who are already working actively in Cyber Security field. This is the reason why CST is split into two courses, while both courses generally follow the same topics, Cyber Security Technologies 2 is designed for students who have some background in information technology, i.e. the student has programming, system administration or information security background. Students in CST I are introduced to topics related to the fundamentals of networking, Information security and cyber security. Students learn about the different types of Cyber Security Technologies, during that process they will understand when and where should these tools be utilised and understand how to configure and deploy specific tools to their own environment. In the second course the students are expected to have some knowledge on these domains and the course covers more advanced aspects of these topics. This enables the course to go more in-depth about the defence technologies taught throughout the course. There can be cases where a student from the advanced course feels that on some topics, he or she does not have enough knowledge then they are free to attend the beginner's course as a listener to learn about the topic from the ground up. Like in the beginners' course the students will learn the theory and then put that learned theory in use in their own lab environment. In order to differentiate what course a student can take there is a skill assessment exam in the first lecture that all the students have to take. If a student's score is above the threshold, he or she can declare the advanced course. However, students are free to attend the lectures of both courses if they are eager to learn more about the subject.

### 2.3.1 Initial Assessment

In order to determine what course the student should take there has to be a way to rate the knowledge and skill set of each individual. Since both courses are focused on technical skills the instructor wants to understand the technical skill set of each individual student. All the students have to attempt to finish a Rangeforce module called the "WASE Assessment". The WASE assessment is a long (180 minutes) and complex virtual lab where the individual has to investigate a website and try to take it back from the hackers. In order to accomplish that the person doing the test has to know different types of web application vulnerabilities and how to exploit them. The types of vulnerabilities are presented in Figure 2 [14].

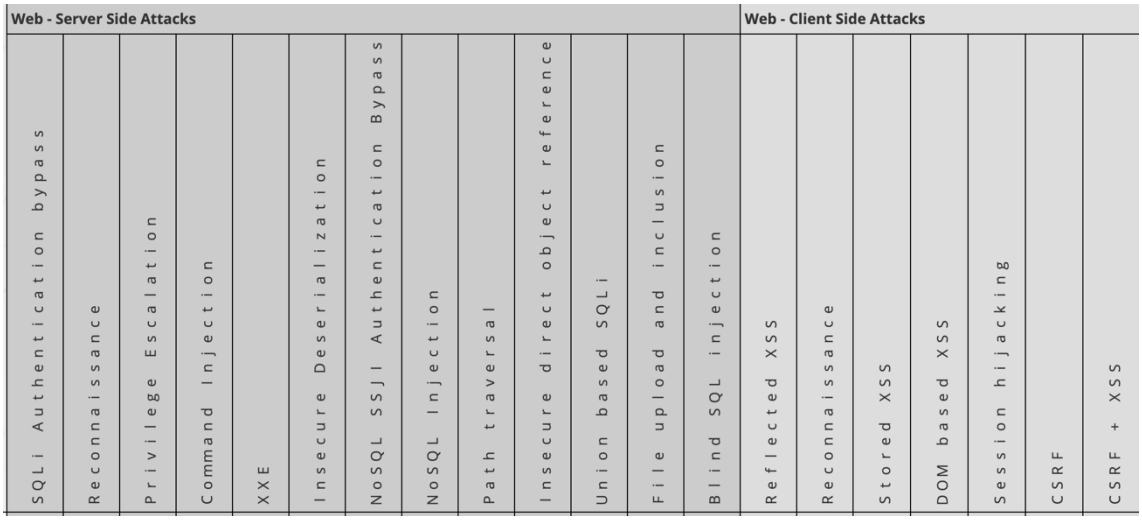


Figure 2. Rangeforce WASE assessment objectives. [14]

As this is a complex and long assessment the expectation was not to see who is able to finish the assessment but rather see how far students could reach within 3 hours. The minimum threshold to get accepted into the advanced class was set to 30 000 points.

### 2.3.2 Course Assignments

The instructors feel that it is important to let the students experiment with the theory that they had been taught in the class. This is why the courses do not have a final exam, but rather challenge the students along the way with different assignments. In order to complete the course, the students have to complete a set of assignments given to them. The assignments consist of group work, home labs, individual assignments, discussions and Rangeforce tests. Groups are self-formed by students during the first class and they will complete all group works with the same group. Recommended group sizes are 4-6

students. Students receive assignments on a weekly basis and generally are given a week to complete them. Assignments assigned throughout the semester are in Table 2.

Table 2 - CST 1 & CST 2 Assignments.

| **Assignment** | **CST 1** | **CST 2** |
|---|---|---|
| Individual Assignment | Malware Lab | Malware Lab |
| Individual Assignment | Quiz | Vulnerability Testing |
| Group Work 1 | Security Principles | Company X – Part 1 |
| Group Work 2 | Information Gathering and Vulnerability Testing | Company X – Part 2 |
| Group Work 3 | Authentication and Access Control | Company X – Part 3 |
| Group Work 4 | Logging and Log Analysis | Company X – Part 4 |
| RangeForce Test | SOC – Security Compromised | SOC – Security Compromised |
| Group Work 5 | Certificates and Public Key Cryptography | Company X – Part 5 |
| Group Work 6 | Risk Management | Company X – Part 6 |

The first assignment is identical for both courses, students are expected to accomplish this alone and at home. Students have to analyze 3 malware samples, answer the questions regarding the malware and create a lab report. The second assignment for CST 1 is a quiz that they answer at home via Moodle. It is a discussion-based quiz, to understand how well the students can express their thoughts. Vulnerability testing assignment is a home task for CST 2 course, where students have to analyze vulnerable server with security tools and answer questions on the subject matter and create a report about it. Group works are started at the end of the lecture and finished at home. Both Rangeforce tests must be completed during the lecture. Students who are unable to attend that lecture have another opportunity to come and take the test. CST 2 course has a dedicated infrastructure at K-Space, students divide into groups, with up to 4 people per group. Each group receives access to a set of virtual machines that they have to do system hardening, defend and deploy defensive cyber security tools on.

## 2.4 Learning Analytics

The usage of data to improve learning is common in universities. This activity usually happens at small scale and in individual courses, where teachers collect data manually or through analysis of server logs to provide individual students with feedback on which exam questions or what learning activities cause learners confusion. This data can also be used to provide feedback to the teachers to measure students' performance on various tests or assignments. This type of data use is helpful to both faculty members and students however, it fails to take advantage of a systematic approach to analytics [16].

The author of the thesis has chosen to use learning analytics approach to answer the research questions because learning analytics defines a set of techniques and algorithms that are used in the learning-related domain [17].

As pointed out by V. Švábenský et al [4], the cybersecurity education as a research field would be improved when employing learning analytics and educational data mining to better understand learners. The predictive analytics is considered important part of this research area [18].

Relevant research includes identifying students at risk and predictive analytics---several methods for student modeling have been developed, however methods have not been widely researched in cyber security (e.g., using log data to predict course grade [19], predicting team proficiency [20]). Ideas such as assigning specific exercise types, aims, and participants to each level of preparedness, could be further validated and can lead to improvement of learners' capabilities and reduce resource misuse (i.e., too complex exercise offered to novices) [21].

Learning analytics has two components techniques and applications [16]. Techniques are the specific algorithms and models used for conducting the analytics and applications involve the way how techniques can be used to impact and improve learning and teaching. For example, an algorithm that provides recommendations of additional course content for learners can be classified as a technique. A technique, such as prediction of learner risk for dropout, can then lead to an application, such as personalization of learning content to reflect learners' comfort with the subject area. The distinction between technique and an application is not absolute but instead reflects the focus of researchers. Commonly used analytics techniques are presented in Figure 2 [16].

| LA Approach | Examples |
|---|---|
| Techniques | |
| Modeling | Attention metadata |
| | Learner modeling |
| | Behavior modeling |
| | User profile development |
| Relationship mining | Discourse analysis |
| | Sentiment analysis |
| | A/B testing |
| | Neural networks |
| Knowledge domain modeling | Natural language processing |
| | Ontology development |
| | Assessment (matching user knowledge with knowledge domain) |
| Applications | |
| Trend analysis and prediction | Early warning, risk identification |
| | Measuring impact of interventions |
| | Changes in learner behavior, course discussions, identification of error propagation |
| Personalization/adaptive learning | Recommendations: Content and social connections |
| | Adaptive content provision to learners |
| | Attention metadata |
| Structural analysis | Social network analysis |
| | Latent semantic analysis |
| | Information flow analysis |

Figure 2. Learning Analytics Techniques and Applications. [16]

The author follows the modeling technique using research model as depicted also in Figure 3, when conducting this research.

# 3 Related Work

This chapter reviews the existing body of knowledge regarding the use of predictive modelling at university programs and the admission process. To find related work the author used Google Scholar and the following keywords to identify relevant papers: university admission and Cyber()security, university admission and STEM /computer science, student (learning) success prediction and Cyber(), student (learning) success prediction and STEM /computer science.

## 3.1 Predictive Models at University Programs

Predictive models have been used in the university admission process before [22], there are several examples of predictive models [22]. There are several attempts to predict students' performance in Science, Technology, Engineering and Mathematics (STEM) disciplines, such as [23], [24] or [25] who also provides related models overview.

Kabra et al. [23] use data mining and decision tree algorithms to predict students' performance using engineering student's past performance data. The goal is to enable identification of students who are likely to fail in advance, this would allow the teacher to adjust and provide appropriate inputs to the students. The paper concludes that using decision tree algorithm can predict students' performance, in addition, the paper mentions that the prediction model could be improved if the collected dataset would include students' current performance (e.g. attendance, test marks etc.) [23].

The literature on learning analytics is replete with studies on the use of data to predict student performance [25]. However, most of those models assume previous knowledge of past performance or are mainly based on demographic data that is not necessarily available in the admission process or during the first study semester. There are attempts, such as [26], aiming to build a prediction model without legacy data (e.g., using the first assignment submission rate as predictor). While there is a lot of related work to work with, the existing body of knowledge does not address the question whether technical lab or gamified simulations can be used as a relevant predictor for future learning success.

Tests/assessments are seen as "means" to early distinguish students who are likely to achieve high or drop out [27]. In addition, such studies have not been done in Cyber Security domain, with one exception being TalTech [8] which describes how the admission process works but does not validate if the model works. This thesis builds upon this previous work and explores this further by aiming to validate that gamified/simulated hands-on test can be used as predictive assessments, as these combine both technical knowledge and practical problem-solving in the simulated learning environment.

## 3.2 Building Prediction Models in STEM

A mathematical model generally consists of a set of mathematical formulas that describe the quantitative (i.e., numerical) relationships between dependent variables (i.e., outputs) and independent variables (i.e., inputs, or predictor variables).

Huang describes that if the instructor of a course wants to predict the average academic performance of his/her class, the instructor should choose the multiple linear regression model. For example, using students' cumulative GPA as the predictor variable. Adding more variables to the situation does not necessarily improve the prediction accuracy. [28]

The model is validated if it makes accurate predictions (i.e., the error between predicted and actual values is within a predefined small range) [28]. In engineering education, [28] describes the four types of mathematical models - the multiple linear regression model, the multilayer perceptron network model, the radial basis function network model, and the support vector machine model - as most widely employed in engineering research and engineering education research. The inputs (i.e., predictor variables) of the models include the student's cumulative GPA, grades earned in four pre-requisite courses (statistics, calculus I, calculus II, and physics), and scores on three dynamics mid-term exams (i.e., the exams given to students during the semester and before the final exam). The output of the models is students' scores on the dynamic's final comprehensive exam [28].

Huang concludes that when predicting the academic performance of individual students, the instructor should use the support vector machine model with the first six predictor variables as the inputs of the model (student's cumulative GPA, grades earned in four pre-requisite courses, and scores on three dynamics mid-term exams), because this

combination increases the percentage of accurate predictions and most importantly, allows sufficient time for the instructor to implement subsequent educational interventions to improve student learning [28]. Other methods are proposed in the literature, such as survival analysis and compared their model with Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB) and Adaboost (AB) models [24], [29]).

Several methods for student modelling have been developed (e.g., Knowledge Tracing, Performance Factors Analysis) [30], however, such methods have not been widely researched in cybersecurity (only a few examples such as [19] using log data to predict course grade, [20] predicting team proficiency). Ideas such as [21] assigning specific exercise types, aims, and participants to each level of preparedness, could be further validated and can lead to improvement of learners' capabilities and reduce resource misuse (i.e., too complex exercise offered to novices).

The analytical metrics include the learners' individual characteristics, such as socio-demographic information, personal preferences and interests, responses to standardized inventories, skills and competencies, prior knowledge and academic performance, as well as institutional transcript data [31]. However, many of such metrics may not necessarily be available and admission decisions need to be made with limited data. Also, the existing work does not address the question of whether gamified hands-on exercises can be used as a relevant predictor for future learning success.

In the cyber security field, [32] describes a Computer Science freshman recruiting tool that provides an eight-hour cyber training and competition framework designed to be attended by Computer Science candidates. However, this approach is not scalable in case of international admissions as assumes attendance and significant time commitment from existing faculty and students.

Campbell et al. [33] propose a model for predicting cybersecurity aptitude beyond a general-intelligence approach. They suggest that tasks, work roles, and people can be represented along the same set of axes to match job requirements to personal attributes. These constructs can then be used to create assessments of the potential for cybersecurity applicants, including the Cyber Aptitude and Talent Assessment as proposed by Campbell. However, the challenge is that the applicants are still exploring different career

paths and the admissions process should allow such flexibility, and also accept different student profiles. Also, such aptitude tests do not reveal technical base-level skills [33].

## 3.3 Learning Analytics and Cyber Security Education

Learning analytics aims to provide evidence-based approach to evaluate learning impact. [16] However, the use of learning analytics in cyber security education, including predictive analytics, is still in the early stages. Cyber security as a field lacks a common agreement how learning analytics and predictive analytics should be applied to be used as skills assessment for cyber security students.

Svabensky et al. [4] suggests researchers to consider using educational data mining and learning analytics approaches to better understand the learning process of students. Researchers should read exemplary papers and perform thorough evaluations when sharing their data sets. The community would benefit from rigorous reports about methods that researchers use, as there are papers where the description of the methods used was incomplete or unclear, making it difficult to follow the same model [4].

This research paper aims to provide a method that other researchers can also use and in addition describe the dataset used - as this is seen as a gap in cyber security educational research.

# 4 Research Method

In this chapter research model used to conduct the research is explained and each step of the model is described. Since the research handles personally identifiable information data of students, concerns about ethics and privacy are also discussed in this part.

## 4.1 Research Model

Learning Analytics Model described in Figure 3 is used as a guideline to conduct the research. Learning analytics model detailed below introduces a systematic approach to analytics. The model includes seven components: collection, storage, data cleaning, integration, analysis, representation and visualization, and action [16].

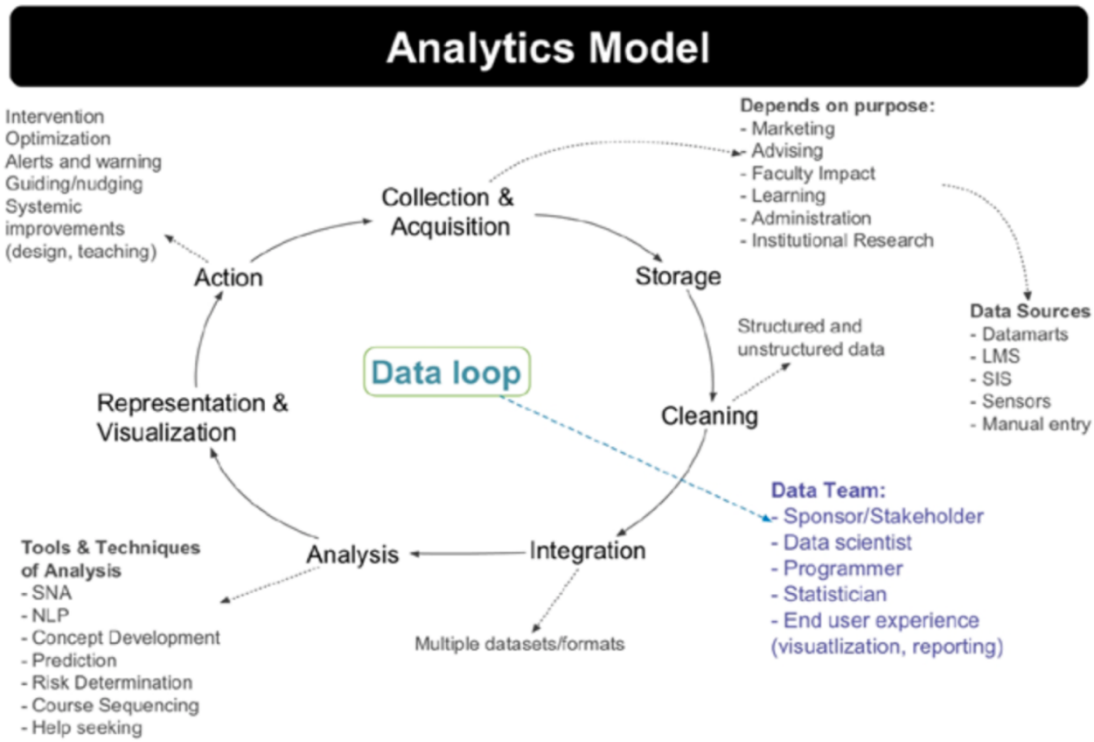


Figure 3. Learning Analytics Model. [16]

### 4.1.1 Collection & Acquisition

The analysis requires data from sources that reflect the learning process. Ideally, data that is captured as learners are engaged in learning.

The collected dataset consists of multiple sources:

- Admission Interview Scores
- Admission Technical Assessment
- Skill Exam
- Group Works
- Individual Assignments
- Rangeforce Tests

Firstly, the data from the online technical assessment are collected from the admission board. This data includes metrics such as labs completed, completion time, percentage of lab tasks completed, etc.

Secondly, the skill test results are collected, this data includes metrics such as completion time, percentage of tasks completed and score.

Then all the tasks that are completed throughout both courses were collected, these included group works, home labs, individual assignments and Rangeforce Tests.

Data collected from these data sources is sufficient to answer the research questions. In the scope of the research are the results of 60 students. The collected data is described in Table 3.

Table 3. Data overview.

| **Field Name** | **Value** | **Unique data values** | **Description** |
|---|---|---|---|
| Student Name | Firstname. Lastname | 60 | Used an identifier to integrate data together from different sources. Once each student receives an ID number, this field is removed. |
| Student ID | Student0X | 60 | When conducting the research each student will be identified with an ID number, starting from 001. |
| Course | Advanced<br>Beginner<br>Advanced/Beginner | 3 | There are three unique values, students who are eligible for CST 2 course are classified as advanced students, students eligible for CST 1 are classified as beginners, people who were eligible for CST 2 but chose to take both courses are classified as advanced/beginners |
| Admission Interview Score | 50 - 100 | 60 | Admission interview results, minimum threshold to accept candidates into the program is 50 points. |
| Admission Online Technical Assessment Score | 0 - 400 | 59 | Admission Technical assignment results - there are 4 different labs, each worth up to 100 points. |
| Assignment 1 Lab Results | 0 - 100 | 60 | The first assignment is identical for both courses, students are expected to accomplish this alone and at home. Students have to analyze 3 malware samples, answer the questions regarding the malware and create a lab report about it.<br>As both courses have to complete this lab, this can be used to measure student performance, regardless of what course the student is enrolled. |
| CST 1 Individual Assignments | 0 - 400 | 52 | Aggregated results of all individual assignments for CST 1 |
| CST 1 Group Works | 0 - 600 | 52 | Aggregated results of all group assignments for CST 1 |

| | | | |
|---|---|---|---|
| CST 1 Course Total | 0 - 1000 | 52 | Aggregated results of all assignments for CST 1 |
| CST 2 Individual Assignments | 0 - 200 | 11 | Aggregated results of all individual assignments for CST 2 |
| CST 2 Group Works | 0 - 300 | 11 | Aggregated results of all group assignments for CST 2 |
| CST 2 Course Total | 0 - 500 | 11 | Aggregated results of all assignments for CST 2 |
| WASE Assessment Score | 0 - 140 000 | 60 | Every student who wants to take either CST 1 or CST 2 has to complete WASE assessment in order to be assigned to the proper course.<br>As every student has completed this assessment, this data can be used to measure and compare student performance. |
| SOC – System compromised Progress | 0 - 100 | 60 | Rangeforce assessment, that all the students have to complete in class, regardless of the course the student is enrolled in. |
| SOC – System compromised Duration | 0 - 5h | 60 | All the students are encouraged to finish the lab, so in this case, measuring progress is redundant as most of the students will get 100% score of the lab, this is why more interesting factor here is time, How far does it take the student to complete the assessment. This can be used to compare students' understanding of the issue at hand and how fast they can solve it. |

### 4.1.2 Storage

Both security and privacy aspects have to be taken into account when storing the research data. Research data of this thesis is stored on TalTech's Sharepoint server, access to the research material is restricted to only the author of the thesis and select few TalTech personnel and TalTech researchers directly associated with this study and the admission process. Only data that is necessary for conducting the research is stored, all the unnecessary info is removed from the server.

### 4.1.3 Cleaning

In this phase, the author converts the unstructured data to structured data. This includes removing irrelevant information from different data sources. In addition, before analyzing data, it is essential that it does not contain any errors, in this stage the author spends time to verify that the values in the data set are correct. Furthermore, any PII data is also removed from the dataset and everything is pseudonymized.

There are a couple of students who take both classes, in such case the duplicate data is removed. In addition, there are students who are removed from the dataset as they are out of scope. There are 3 different scenarios why students may be out of scope:

1. The student is not part of Cyber Security Master's Program (he or she has selected this subject as free select subject)

2. Student enrolled in Moodle but did not declare the course in study information system (he or she is registered in Moodle as a student, but has actually decided not to take the subject)

3. The student is part of Cyber Security Master's Program but is not a 1st-year student (in this research only 1st-year students are in scope)

Four students match one of the aforementioned criteria and thus are out of scope and removed from the research.

### 4.1.4 Integration

The data from all the different sources are integrated together into one set of data in order to identify relevant metrics. At this stage, the author introduces newly aggregated fields into the data set named "Ranks". This data is used to rank the students in a unified format and based on these values the correlation methods are used to answer the research questions. In the table below are represented the values that are calculated by the author, see Table 4.

Table 4 - Ranking system.

| Field Name | Calculation Method | Description |
|---|---|---|
| Admission Interview Rank | RANK(SUM(Admission Interview);desc) | Ranks the students based on admission interview results |
| Admission Technical Assessment Rank | RANK(SUM(Admission Technical);desc) | Ranks the students based on admission technical assessment results |
| Admission Rank | SUM(Admission Interview rank + admission technical rank)<br>RANK(Admission rank sum) | The combined rank of admission interview and admission technical lab |
| WASE Assessment Rank | RANK(SUM(WASE Assessment);desc) | Ranks the students based on WASE assessment results |
| Lab 1 Rank | RANK(SUM(Lab 1);desc) | Ranks the students based on lab 1 results |
| SOC – System compromised Rank | RANK(SUM(System compromised);asc) | Ranks the students based on Rangeforce system compromised results |
| Student Rank Without WASE Assessment | RANK(SUM(Lab 1);desc)<br>+ RANK(SUM(System compromised);asc) | Ranks the students based on lab 1 and system compromised |
| Student Final Rank | RANK(SUM(WASE Assessment);desc) + RANK(SUM(Lab 1);desc)<br>+ RANK(SUM(System compromised);asc) | The combined rank of WASE assessment, Lab 1 and system compromised rank |

### 4.1.5 Analysis

Correlation analysis is used to describe the strength and direction of the linear relationship between two variables. In this thesis, Pearson's correlation coefficient and Kendell Tau measure are used to analyze relationships. Pearson's correlation coefficient test is used to test and measure the statistical relationship between two variables. The test is known as the best method to measure association between variables of interest as it is based on

covariance method. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship [34].

Kendall's Tau is a non-parametric measure of relationships between columns of ranked data. The Tau correlation coefficient returns a value of 0 which means there is no relationship between the data or 1 there is a relationship between the data [35].

Since there is no comparable research, the author uses conventions to decide the effect of the predictor. To describe, if effects have a relevant magnitude, Cohens' d effect size measure is used to describe the strength of a phenomenon [36].

For the calculation of the correlation relationship, the author used open-source software named JASP.

### 4.1.6 Representation & Visualization

Usually during this phase results of the analysis are presented and visualised. The author uses the visualization as part of answering the research questions and visualize important data to identify patterns and outliers.

### 4.1.7 Action

After the conclusion of data analysis, the research questions are answered and relevant action can be taken based on that information. For example, if the research paper validates that Rangeforce Technical assessment is a valid predictor for student performance the author can make a suggestion that the assessment should be a mandatory step part of the admission process. Then the admission board can take this into consideration and implement this into the process.

## 4.2 Ethics and Privacy

Privacy and data ownership concerns are not unique problems to the analytics field, any type of online or digital interaction produces a data trail [16].

Collection, processing and analysis of data have faced an abundance of ethical breaches and constraints in the past. Revealing learners' personal information, activities and other major aspects of their personal lives that make it possible to identify individuals. The

process of de-identification of data can reduce the risk of inadvertent disclosure of learners' identities. De-identification techniques include methods like anonymization, masking, blurring and perturbation. The last step of de-identification is to link the anonymized data with a unique descriptor, so that the data may be examined by the researched [37].

In this paper, all the data collected has been anonymized and linked with a unique descriptor to ensure that ethics and privacy of individuals are not violated.

# 5 Analysis of Data

In this chapter the data is analysed to answer the research questions. The chapter is divided into two subsections. The first subsection presents the descriptive statistics report about the data important to answer the research questions. In the second section, visualisation and correlation techniques are used to answer the research questions.

## 5.1 Descriptive Statistics

The main function of descriptive statistics is to summarize large chunks of data into information that is meaningful. In the dataset under research, there are results of 60 different students. The author presents a statistic report about the following data:

1. Admission Labs Results
2. Admission Interview Results
3. WASE Assessment
4. Lab 1 Results
5. Rangeforce System Compromised Points
6. Rangeforce System Compromised Time
7. CST 1 Individual Assignments
8. CST 1 Group Assignments
9. CST 1 Course Total Score
10. CST 2 Individual Assignments
11. CST 2 Group Assignments
12. CST 2 Course Total Score

The descriptive statistic is presented in Table 5.

Table 5 - Descriptive statistics.

| | **Mean** | **Median** | **Min** | **Max** | **Standard Deviation** | **Range** |
|---|---|---|---|---|---|---|
| Admission Labs Results | 280.2 | 293.0 | 0 | 400 | 117.8 | 400 |
| Admission Interview Results | 73.90 | 77.25 | 50 | 97.5 | 12.09 | 47.5 |
| WASE Assessment | 32 000 | 20 000 | 0 | 150 000 | 40 497 | 150 000 |
| Lab 1 | 89.5 | 100 | 0 | 100 | 26.32 | 100 |
| Rangeforce System Compromised Points | 85.05 | 100 | 0 | 100 | 32.7 | 100 |
| Rangeforce System Compromised Time | 1.9 | 2.07 | 0 | 4,42 | 0.98 | 4,42 |
| CST 1 Individual Assignments | 249.1 | 268.3 | 0 | 310 | 63.15 | 310 |
| CST 1 Group Assignments | 354.13 | 380 | 0 | 400 | 91.76 | 400 |
| CST 1 Course Total | 687.9 | 739.05 | 81.25 | 845 | 160.74 | 763.75 |
| CST 2 Individual Assignments | 175 | 200 | 0 | 200 | 60.2 | 200 |
| CST 2 Group Assignments | 189.1 | 190 | 170 | 200 | 10.44 | 30 |
| CST 2 Course Total | 473.2 | 490 | 300 | 590 | 68.3 | 290 |

The descriptive statistics intend to show transparency into the collected data and allow for comparisons with other (also future) studies/samples. In the presented statistics report the author included mean, median, min, max, standard deviation and range. This is the minimum information needed to get an idea of what the distribution of data set looks like.

## 5.2 Research Questions

### 5.2.1 How do interview and hands-on technical assessment labs components of the admission and selection process and their results predict student performance?

To start answering RQ1, the dataset is first explored to detect any patterns emerging in visualizations to analyse the significance of labs versus interviews. The admission data is presented in scatter plot graph with admission lab score on the Y-axis and Interview score on the X-axis in Figure 4. In order to get accepted to the Cyber Security Master's program the candidate must receive at least 50 points from the interview and that means by definition thus an accepted student could not have less than 50 points, this is why X axis starts from 50 points.

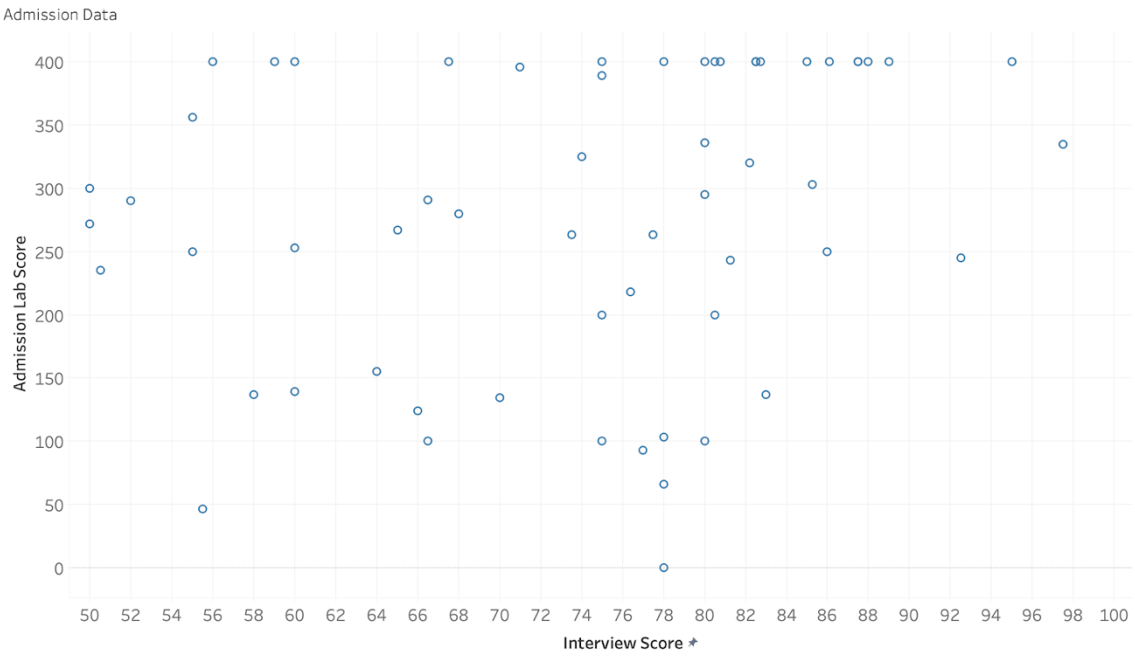


Figure 4. Admission Results Scatterplot.

In general, the higher the candidates interview score the higher the lab score is as well. However, this is not a rule, there are many exceptions. There are students who received 50-60 points during the interview but managed to get maximum points from the technical assessment. In contrast, there are also students who received higher scores from interview but received 50-100 points from the technical assessment. Another example is of a student who received 92.5 points from the interview but only 250 points from the technical assessment. This shows that the combination of interview and technical assessment can

be valuable, as a person who might not have technical skills can still get higher score in the interview and vice versa.

In order to answer the research question Pearson's one-sided correlation testing method is used because the author has directed hypothesis that the admission results and CST-performance are positively related. The following four variables are added to the correlation matrix:

- Course performance
- Admission interview rank
- Admission technical assessment rank
- Admission rank

The results of the correlation matrix are in Table 6.

Table 6. Admission results Pearson's one-sided correlation test.

| **Course Performance** | **Admission Interview Rank** | **Admission Lab Rank** | **Admission Rank** |
|---|---|---|---|
| Pearson's r | 0.378** | 0.432** | 0.492*** |
| p-value | 0.005 | 0.001 | < .001 |

From the output given above, there is a strong positive correlation between the input and all three variables, while the strongest positive correlation is between Course performance and Admission Rank where the correlation coefficient is 0.492. This is because of the fact that the Admission Rank is based on the two variables "Admission Interview Rank" and "Admission Lab Rank". The relationship between the input and interview variable is the weakest, but the interview consists of different variables in its own that cannot always be quantified compared to practical technical assignment. While the lab scenarios are all the same for the students, the interview can vary with each candidate and the interviewees' mood and nervousness can also affect the score, with this being said, it is still considered a strong correlation.

Multiple regression is not just one technique but a family of techniques that can be used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors (usually continuous). Multiple regression is based on

correlation but allows a more sophisticated exploration of the interrelationship among a set of variables. Stepwise regression is used in order to explore the relationship between Course performance and admission results.

Admission interview results and admission technical results are added to the model as independent variables and course performance as the dependent variable. In the null model the components used are course performance as the dependent variable and admission interview rank as covariate. Model one includes admission interview rank and admission technical rank as the covariates and course performance as dependent variable. The linear regression model summary results are in Table 7.

Table 7. Admission results linear regression model summary.

| Model | R | R2 | Adjusted R2 | RMSE | R2 Change | F Change | df1 | df2 | p |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.378 | 0.143 | 0.126 | 14.448 | 0.143 | 8.507 | 1 | 51 | 0.005 |
| 1 | 0.495 | 0.245 | 0.214 | 13.700 | 0.102 | 6.727 | 1 | 50 | 0.012 |

When looking at the interview as the predictor it significantly (.005) predicts the student performance and explains .126=12.6% of the student performance variance. This means that interview in its own is already a valid predictor for student performance score, but when adding the virtual lab, the R square almost doubles from 12.6 to 21.5. This increase is significant (p = .012) - which means that interview and lab are complementary methods that both predict in different aspects the student performance results.

The results for analysis of variance (ANOVA) test are presented in Table 8.

Table 8. Admission results ANOVA test.

| Model | | Sum of Squares | df | Mean Square | F | p |
|---|---|---|---|---|---|---|
| 0 | Regression | 1775.822 | 1 | 1775.822 | 8.507 | 0.005 |
| | Residual | 10646.291 | 51 | 208.751 | | |
| | Total | 12422.113 | 52 | | | |
| 1 | Regression | 3038.279 | 2 | 1519.139 | 8.094 | <. 001 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Residual | 9383.834 | 50 | 187.677 | | |
| | Total | 12422.113 | 52 | | | |

Relationships between the predictor variable and response is shown in Table 9.

Table 9. Admission results coefficients.

| **Model** | | **Unstandardized** | **Standard Error** | **Standardized** | **t** | **p** |
|---|---|---|---|---|---|---|
| 0 | (Intercept) | 17.023 | 3.899 | | 4.366 | <.001 |
| | Admission Interview Rank | 0.340 | 0.117 | 0.378 | 2.917 | 0.0005 |
| 1 | (Intercept) | 13.639 | 3.920 | | 3.479 | 0.0001 |
| | Admission Interview Rank | 0.231 | 0.118 | 0.257 | 1.955 | 0.056 |
| | Admission Technical Assessment Rank | 0.246 | 0.095 | 0.341 | 2.594 | 0.012 |

The standardized betas in Table 9 are the relative weight of the predictors. When comparing the standardized betas, they are all rather weak, but when combining the predictors as it is done in reality the overall model is highly significant. The predictors' scores are relatively highly correlated, which means that while the interview scores and technical skills results predict student performance from different aspects, they also have common factors. For example, attitude, eagerness and interest are checked in the interview process but are also relevant for the technical performance.

### 5.2.2 Do the Web Application Security (WASE) assessment and its results predict student performance?

Firstly, analyzing visualization of WASE assessment, Figure 5, it is visible that people who have completed WASE assessment in IT College and receive maximum points of

150 000 are automatically accepted into the advanced class. Five people who received 150 000 points from the WASE assessment are all former IT College students. Out of these five students two students wanted to take both courses and enrolled in CST 1 as well. There are 6 students who reached 30 000 points but felt that they would rather start with the CST 1 instead. 29 students received 20 000 points and 13 people did not receive any points from the assessment. Eleven people chose CST 2, while three of them also declared CST 1 as well. It must be noted that there are two students from CST 1 that were not eligible to sign up for CST 2 but chose to attend the lectures anyway as they are eager to learn more about the subject.

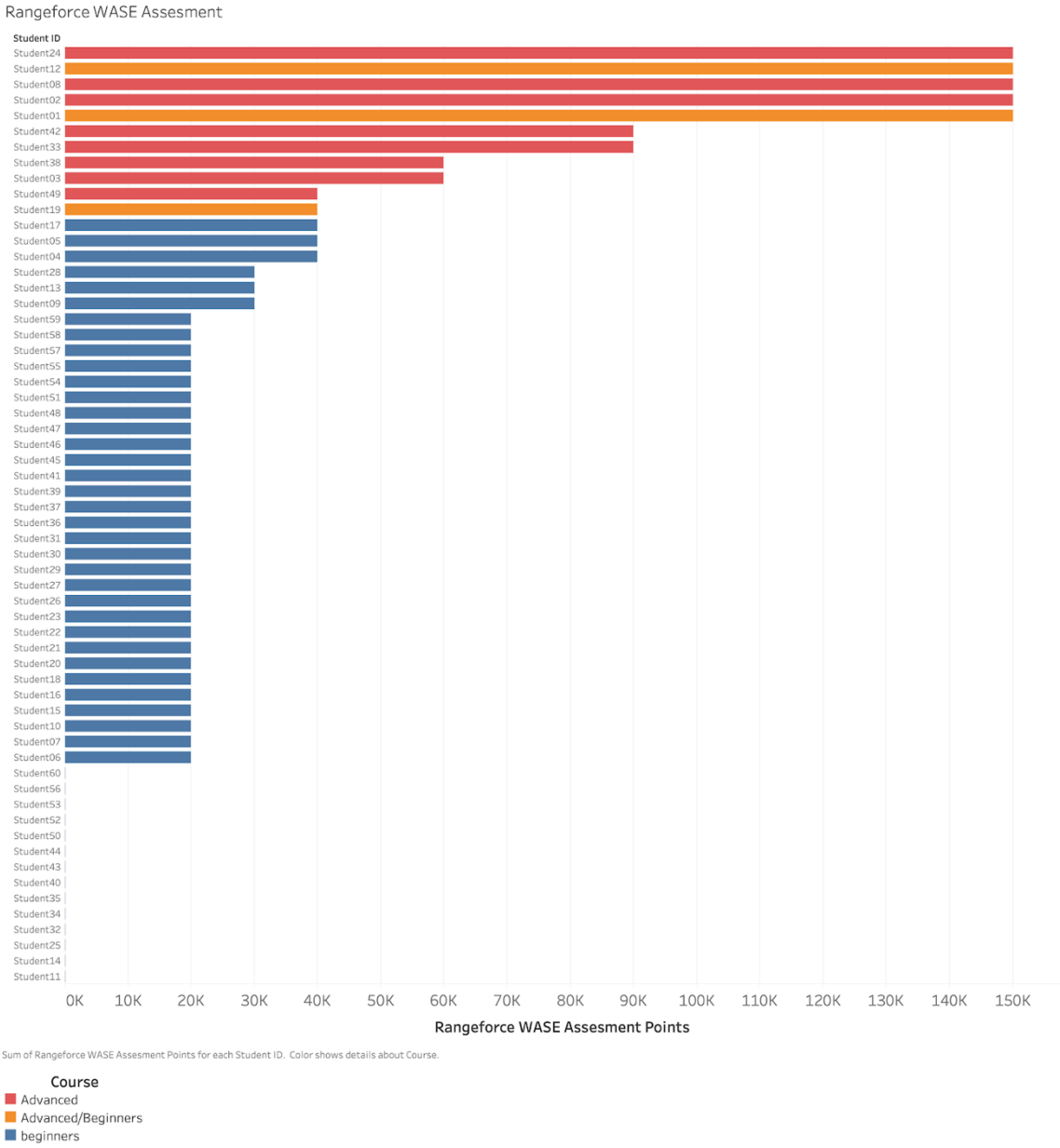


Figure 5. WASE assessment results.

To answer the research question, whether the WASE assessment and its results predict student performance and in turn assignee the students to the correct Cyber Security Technologies course. The author uses Kendall's Tau B testing method. With this testing method, we can check if the WASE assessment has a relationship with the course performance. The Kendall's Tau B test results are in Table 10.

Table 10. WASE assessment Kendell Tau result.

| **Course Performance** | **WASE Assessment** |
|---|---|
| Kendall's tau B | 0.502** |
| p-value | < .001 |

From the output given above, there is a strong positive correlation between WASE Assessment and course performance. This means that WASE assessment is a valid predictor for students' performance.

### 5.2.3 Is the WASE assessment necessary, if the admission technical assessment predicts student performance?

In order to answer the research question Pearson's one-sided correlation testing method is used because the author has directed the hypothesis that WASE assessment and Admission technical results are positively related with CST performance results. In the "course performance" variable WASE assessment results are part of the calculation, but, as we are comparing course results to WASE assessment results, the author is using "course performance without WASE" variable instead. Three variables are added to the correlation matrix:

- Course performance without WASE
- Admission technical rank
- Admission technical assessment rank

The results of the correlation table are in Table 11.

Table 11. WASE & technical assessment Pearson's one-sided correlation test.

| **Course Performance Without WASE** | **Admission Technical Rank** | **WASE Assessment Rank** |
|---|---|---|
| Pearson's r | 0.322* | 0.548*** |
| p-value | 0.019 | .001 |

From the output given above, there is a significant positive correlation between the input and admission technical rank, however, the correlation between Course performance and WASE assessment rank is a lot stronger, the correlation coefficient is 0.548.

Stepwise regression is used to further explore the relationship between Course performance, admission technical rank and WASE Assessment rank. In this case course performance without WASE assessment is the dependent variable and "Admission technical rank" and "WASE assessment rank" are the predictor variables. The linear regression model summary results are in Table 12.

Table 12. WASE & technical assessment model summary.

| **Model** | **R** | **R2** | **Adjusted R2** | **RMSE** | **R2 Change** | **F Change** | **df1** | **df2** | **p** |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.322 | 0.104 | 0.086 | 14.727 | 0.104 | 5.890 | 1 | 51 | 0.019 |
| 1 | 0.551 | 0.304 | 0.276 | 13.106 | 0.200 | 14.401 | 1 | 50 | <.001 |

When looking at the admission technical results as the predictor it significantly (.019) predicts the student performance and explains .086=8.6% of the Student performance variance. This means that interview in its own is already a valid predictor for assigning students into correct CST course. However, when adding the WASE assessment the R square is 3 times higher from 8.6 to 27.6. This increase is significant (p = .001) - which means that the two variables are complementary methods that both predict in different aspects of the student performance results.

The results for ANOVA test are presented in Table 13.

Table 13. WASE & technical assessment ANOVA test results.

| Model | | Sum of Squares | df | Mean Square | F | p |
|---|---|---|---|---|---|---|
| 0 | Regression | 1277.518 | 1 | 1277.518 | 5.890 | 0.019 |
| | Residual | 11061.275 | 51 | 216.888 | | |
| | Total | 12338.792 | 52 | | | |
| 1 | Regression | 3751.013 | 2 | 1875.506 | 10.920 | <. 001 |
| | Residual | 8587.780 | 50 | 171.756 | | |
| | Total | 12338.792 | 52 | | | |

There was a statistically significant difference between groups as determined by ANOVA ($F(2,50) = 10.920$, $p =< .001$)

Relationships between the predictor variable and response is shown in Table 14.

Table 14. WASE & technical assessment coefficients.

| Model | | Unstandar-dized | Standard Error | Standar-dized | t | p |
|---|---|---|---|---|---|---|
| 0 | (Intercept) | 20.719 | 3.236 | | 6.403 | <.001 |
| | Admission technical rank | 0.231 | 0.095 | 0.322 | 2.427 | 0.019 |
| 1 | (Intercept) | 14.346 | 3.334 | | 4.303 | <.001 |
| | Admission Technical Rank | 0.047 | 0.098 | 0.066 | 0.484 | 0.630 |
| | WASE Assessment Rank | 0.543 | 0.143 | 0.516 | 3.795 | <.001 |

The combined weight of the predictors scores are relatively highly correlated, which means that the admission technical assessment and WASE assessment have common factors. The correlation is relatively high, which means that both methods contribute also individual and unique parts that predict the Student Performance score.

### 5.2.4 Is the course of an appropriate difficulty that is neither too easy nor too difficult?

In order to understand whether the courses CST 1 and CST 2 are of an appropriate difficulty to the students and if the home assignments result in a broad distribution of scores that enables the assessors to differentiate between individuals. Histogram graphs are used to analyse the students course total score. Figure 6 shows the distribution of student scores for CST 1.

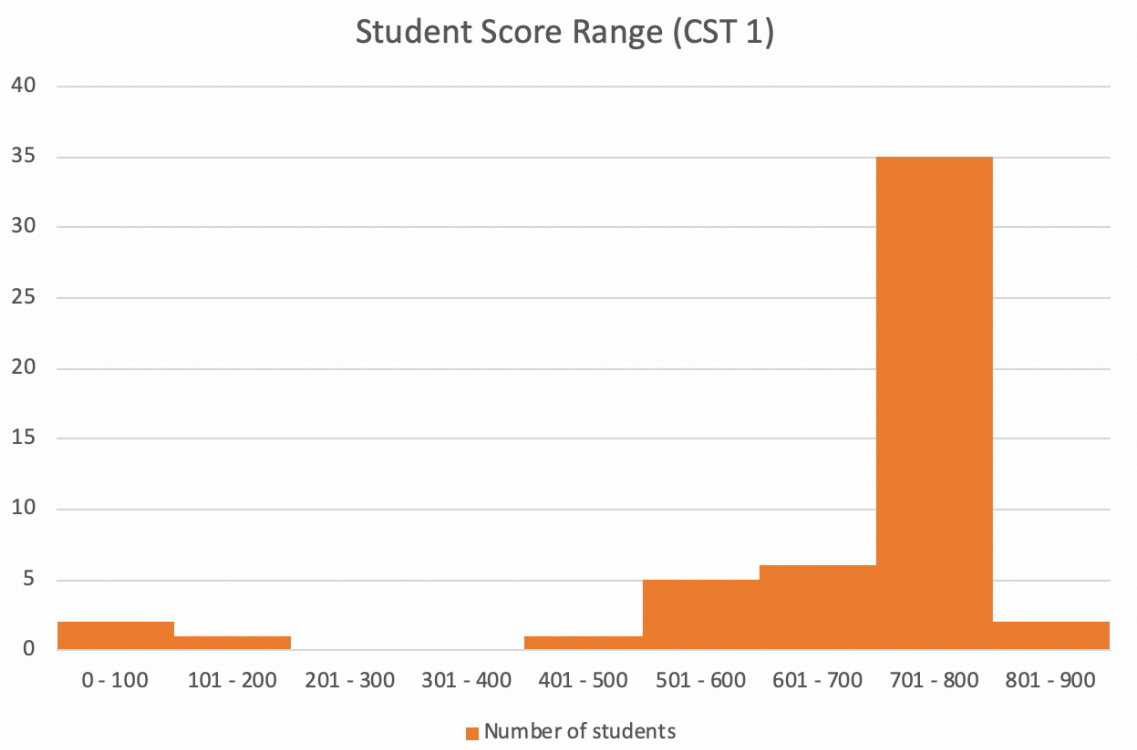


Figure 6. CST 1 student score histogram.

Score distribution in the CST 1 course is not equally distributed. Most of the students who attended the course received a higher amount of points, at first glance this would suggest

that the course assignments are too easy for the students. We also have to take into account the fact that most of the home assignments are group works (6 group assignments versus 3 individual assignments). This means that it is difficult to differentiate between individuals. Overall it is hard to differentiate individual students and answer the research question.

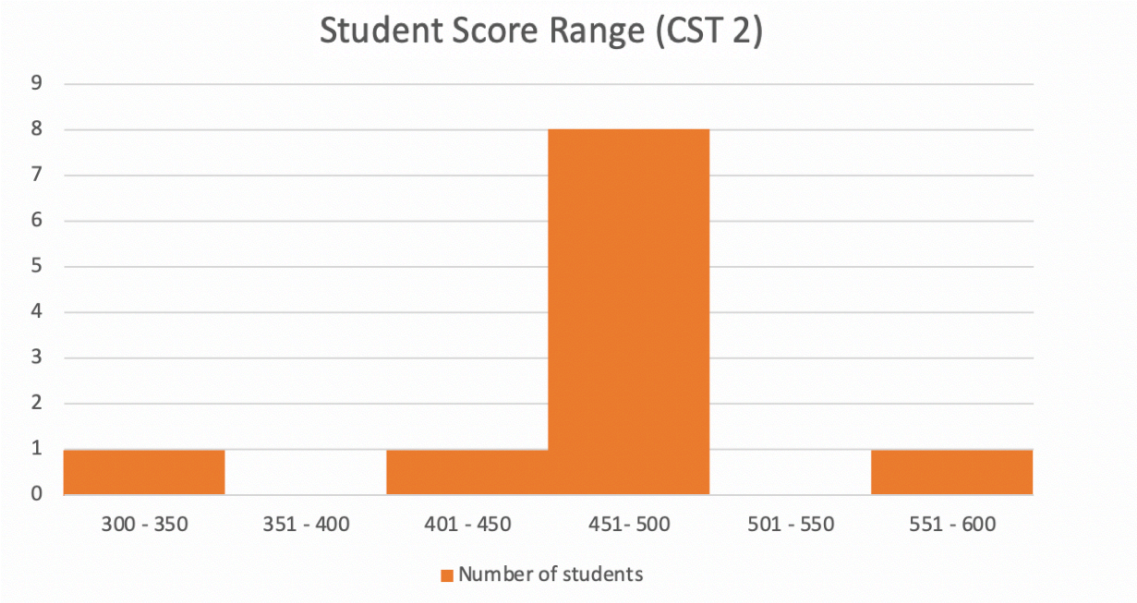


Figure 7. CST 2 student histogram.

Most of the students attending CST 2 received higher amount of points, but only 1 student is at the 551-600 point range, indicating that the course is not that easy, however the pool size for data is relatively small with only 12 results and like in CST 1 we have to take into account the fact that the majority of the points that students receive throughout the course are from group assignments. This means that it is hard to answer the research question because the student pool size is too small and most of the assignments are group works making it hard to differentiate individual student performance.

# 6 Conclusion of Analysis

The admission technical assessment is a valid predictor for student performance score, in addition when including the admission interview results into the predictor model the correlation doubles. The correlation is relatively high, which means that both methods contribute also individual and unique parts that predict the student performance score. They are related to each other (common factors that may be in interview questions on technological skills, knowledge and psychological factors such as motivation, attitude, etc. However, their correlation is not strong, but when both are put into multiple linear regression, the explanatory power of the model is very high for social science method standards. The effect size for admission procedure is large, this means that the admission procedure can predict student performance, thus the first research question is positively validated.

Kendell Tau's correlation method reveals that WASE assessment and course performance have a positive relationship, meaning that the assessment predicts student performance. This also means that the students have been assigned to the correct CST course.

The admission technical assessment has a significant correlation with student performance, and the effect size is intermediate, however, the WASE assessment has a much stronger correlation with student performance and the effect size is large. This means that the technical assessment results could be used to assign students to the course, however, the WASE assessment has a much stronger relationship with student performance, which means that it is a more accurate method to use.

The author is unable to validate whether the CST 1 and CST 2 courses are of an appropriate difficulty that is neither too easy nor too difficult for the students. These results currently point that the assignments are relatively easy, however, most of the home assignments are group works, so it is difficult to analyse individual performance. In addition, the pool size of students should be higher. There are two ways of how this could be improved in the future. Firstly, more individual assignments to differentiate the results and ask for student feedback about the course difficulty. This is valuable information to the course instructors, as they can improve the home assignments for next year.

To conclude, technical skill assessment can be used as a significant predictor to assess potential candidates' skill level and future success in the studies in technical topics. With this being said the author concludes that the interview score is not redundant either. Instead, both of the admission methods are rather complementary to each other, both of the admission procedures address different, but equally relevant aspects of the student performance. To conclude Rangeforce technical skill assessment test is a valid predictor of student performance. The author suggests making the technical skill assessment as a mandatory part of the admission process and use the combination interview score and technical assessment skill test to accept students into TeleTech's cyber security master's program.

# 7 Further Research

There are different opportunities for further research regarding this topic. Firstly, the number of subjects under research currently is 60 students, however, the research would benefit when this number would be increased, it would enable to see if there is a pattern and validate that the admission process can indeed predict students' success.

In addition, this thesis did not take into consideration possible cheating as there was not enough time to look into it in this paper. This would be an interesting topic to solve since the admissions technical assessment is completed online, it leaves room for people to receive outside help or even have somebody else complete the assessment for them.

Furthermore, the research is currently based on the data from one course. The research focused more on the technical skills of the students, but it would be interesting to widen the scope and collect student performance data from other courses as well. This may change the outcome of the research, as some students may be weaker in the technical side of cyber security but, they may be talented on the management side of cyber security.

# References


[1] Strategy Analytics, “Internet of Things Now Numbers 22 Billion Devices But Where Is The Revenue”, https://news.strategyanalytics.com/press-release/iot-ecosystem/strategy-analytics-internet-things-now-numbers-22-billion-devices-where (12.12.2019)

[2] HelpNetSecurity, Number of connected devices reached 22 billion, where is the revenue” [https://www.helpnetsecurity.com/2019/05/23/connected-devices-growth/ (12.12.2019)

[3] EY, “Cybersecurity regained:preparing to face cyber attacks” https://www.ey.com/Publication/vwLUAssets/ey-cybersecurity-regained-preparing-to-face-cyber-attacks/$FILE/ey-cybersecurity-regained-preparing-to-face-cyber-attacks.pdf (12.12.2019)

[4] V.Švábenský, J. Vykopal, P. Čeleda, “What Are Cybersecurity Education Papers About? A Systematic Literature Review of SIGCSE and ITiCSE Conferences.” 2019.

[5] K. Cabaj, D. Domingos, Z. Kotulski, A. Respício. “Cybersecurity education: Evolution of the discipline and analysis of master programs.” 2017.

[6] C.Paulsen, E. McDuffie, W. Newhouse, & P. Toth. “NICE: Creating a cybersecurity workforce and aware public. IEEE Security & Privacy, 10(3), 76- 79”, 2012.

[7] NIST, “Guide for Cybersecurity Event Recovery.” https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-184.pdf (12.12.2019)

[8] K. Maennel; S. Mäses; S. Sütterlin; M. Ernits, O. Maennel. “Using technical cybersecurity exercises in university admissions and skill evaluation. Proceedings of the 14th IFAC/IFIP/IFORS/IEA Symposium on Analysis Design and Evaluation of Human Machine Systems”, HMS : September 16-19, 2019, Tallinn, Estonia, 2019.

[9] TalTech. “Cyber Security”, https://www.ttu.ee/studying/tut_admission/programmes-in-taltech/masters/cyber-security/ (12.12.2019)

[10] Rangeforce, “Introduction”, www.rangeforce.com (12.12.2019)

[11] SAS, “Predictive Modeling Techniques”, https://www.sas.com/ko_kr/insights/analytics/predictive-modeling-techniques.html (12.12.2019)

[12] TalTech, “Cyber Security Specialities” https://www.ttu.ee/studying/tut_admission/programmes-in-taltech/masters/cyber-security/#specialty-12 (12.12.2019)

[13] CrunchBase, “Rangeforce”, https://www.crunchbase.com/organization/vequrity (12.12.2019)

[14] M. Ernits, K. Maennel, S. Mäses, T. Lepik and O. Maennel, “From Simple Scoring Towards a Meaningful Interpretation of Learning in Cybersecurity Exercises”, 15th International Conference on Cyber Warfare and Security March 2020, Norfolk, Virginia [forthcoming].

[15] P. Schaller, M. Schläpfer, “Applied Information Security - A Hands-on Approach”, 2012

[16] G. Siemens, “Learning Analytics: The Emergence of a Discipline”, American Behavioral Scientist, 2013.

[17] D.Song, “Learning Analytics as an Educational Research Approach. International Journal of Multiple Research Approaches”, 10(1):102-111, 2018.

[18] K. Verbert, N. Manouselis, H. Drachsler, E. Duval. “Dataset-driven research to support learning and knowledge analytics”, 2012.

[19] E. Caliskan, U. Tatar, H. Bahsi, R. Ottis, and R. Vaarandi. “Capability Detection and Evaluation Metrics for Cyber Security lab Exercises”, In ICMLG2017 5th International Conference on Management Leadership and Governance. Academic Conferences and publishing limited, 407, 2017.

[20] D. Henshel, G. Deckard, B. Lufkin, N. Buchler, B. Hoffman, P. Rajivan, S. Collman., “Predicting proficiency in cyber defense team exercises. In Military Communications Conference”, MILCOM 2016-2016 IEEE. IEEE, 776–781, 2016.

[21] T. Aoyama, T. Nakano, I. Koshijima, Y. Hashimoto, and K.Watanabe, ” On the Complexity of Cybersecurity Exercises Proportional to Preparedness. Journal of Disaster Research”, Vol 12, 5, 1081, 2017.

[22] J. Campbell, P. DeBlois, and D. Oblinger, "Academic analytics: A new tool for a new era." EDUCAUSE review 42.4: 40, 2017.

[23] R.R. Kabra, R. S. Bichkar, "Performance prediction of engineering students using decision trees." International Journal of computer applications” , 2011.

[24] Y. Chen, A. Johri, and H. Rangwala, "Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early." Proceedings of the 8th International Conference on Learning Analytics and Knowledge. ACM, 2018.

[25] B. Dietz-Uhler, J.E. Hurn, "Using learning analytics to predict (and improve) student success: A faculty perspective." Journal of interactive online learning 12.1 :17-26 , 2013.

[26] M. Hlosta, Z. Zdrahal, and J. Zendulka. "Ouroboros: early identification of at-risk students without models based on legacy data." Proceedings of the Seventh International Learning Analytics & Knowledge Conference. ACM, 2017.

[27] Z. Papamitsiou, et al, "Explaining learning performance using response-time, self-regulation and satisfaction from content: an fsQCA approach." Proceedings of the 8th international conference on learning analytics and knowledge. ACM, 2018.

[28] S. Huang, N. Fang. "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models." Computers & Education 61: 133-145, 2013.

[29] J. Gardner, and C. Brooks, “Evaluating predictive models of student success: Closing the methodological gap.” arXiv preprint arXiv:1801.08494, 2018.

[30] Y. Gong, J.E. Beck, and N.T. Heffernan, "How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis." International Journal of Artificial Intelligence in Education 21.1-2, 27-46, 2011.

[31] C.H. Loh, Y. Sheng, and D. Ifenthaler. "Serious games analytics: Theoretical framework." Serious games analytics. Springer, Cham, 3-29, 2017..

[32] T.A. Augustine, et al. "Cyber competitions as a computer science recruiting tool." Journal of Computing Sciences in Colleges 26.2:14-21, 2010.

[33] S.G. Campbell, P. O’Rourke, M.F. Bunting, "Identifying dimensions of cyber aptitude: the design of the cyber aptitude and talent assessment." Proceedings of the Human Factors

and Ergonomics Society Annual Meeting. Vol. 59. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2015.

[34] Statistics Solutions, “Pearson’s Correlation Coefficient.” https://www.statisticssolutions.com/pearsons-correlation-coefficient/ (12.12.2019)

[35] Statistics How To, “Kendall’s Tau”, https://www.statisticshowto.datasciencecentral.com/kendalls-tau/ (12.12.2019)

[36] W. Lenhard, A. Lenhard. “Calculation of Effect Sizes”, 2016.

[37] M. Khalil, M. Ebner. “De-Identification in Learning Analytics”, 2016.