UNIVERSITY OF TWENTE.

Faculty of Behavioural Management Sciences

Summative Digital Testing in Undergraduate Mathematics.

To what extent can digital testing be included in first year calculus summative exams, for Engineering students?

Alisa J. Lochner
M.Sc. Thesis
January 2019

Examination Committee:
dr. J. T. van der Veen
prof. dr. ir. B. P. Veldkamp

Educational Science and Technology
Faculty of Behavioural Management Sciences
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

MASTER THESIS

Title

SUMMATIVE DIGITAL TESTING IN UNDERGRADUATE MATHEMATICS

Author A. J. LOCHNER

Graduation Committee

1st supervisor DR. J. T. VAN DER VEEN

2nd supervisor PROF. DR. IR. B. P. VELDKAMP

Version: Public

Table of Contents

Acknowledgements

The completion of my Masters in Educational Science and Technology marks the end of a three-year journey of planning, preparing and doing. I would never have come to the Netherlands in the first place, without those that believed in my dream in building a better society with the help of technology through open online Mathematics education that has adaptive formative testing and effective feedback. I would never have come if it were not for those around me that refused to quit when challenges came. They believed in me and my dreams, when I could not. To Joyce Stewart, Belarani Kanjee and my Mother, you still inspire me with your tenacity to this day. To my aunt, Wilna, thank you for your support.

Thank you to those at the Navigators Student Christian Society that gave me a sense of belonging, a sense of family and teaching me the importance of community, relationships and grace. Special mention goes to: Henrike, Margreeth, Wietske, Elisabeth and Ilse. To my (ex-)housemates Remco and Daphne. Thank you for giving me a literal home and always listening. Luiza and Dijana, thank you for being my academic buddies. Dear Allan, my most precious friend. I could write a whole acknowledgement page to you. In short, thank you for carrying this thesis with me, supporting me, praying with/for me, and displaying the heart of Jesus to me.

To my supervisors, Jan van der Veen and Bernard Veldkamp. I knew very little about you when I started my thesis. Soon I realised just how much you both have achieved and know, which shows through your wisdom, advice and ideas. Thank you, especially Jan, for all your support, time, patience and enthusiasm. It really is much appreciated, and it has been both a pleasure and privilege working with you. Thank you to those in the digital testing group for your support and trusting me with your pilot data. Thank you to Tracy for all your help, insight, laughs and for reminding me that about just how beautiful Mathematics is. A thank you to the University of Twente for giving me a bursary so I could complete my Master studies. To all the staff in the EST department, the library, the international office, and student services. Thank you for all you do. A special thank you to Leonie, for listening and all your support and encouragement in difficult times.

To my parents. You always say that all that is needed is a thank-you. You deserve so much more as I would not have been here without your support, encouragement, love and faith in me. Thank you for not giving up, and encouraging me to be strong, whilst also making me laugh. I appreciate all the sacrifices you have made to get me where I am today.

To my saviour and my foundation: Lord Jesus Christ. Thank You for always being beside me and giving me the opportunity and strength to finish this thesis. To You be the power, glory and honour forever and ever. Amen.

Abstract

With increasing student numbers across universities, digital testing serves as a potential solution to the tedious time-intensive marking of exams. This research investigated to what extent digital testing that uses multiple choice and final answer items could assess first year calculus for engineering studies. In order to investigate this, a mixed methods study was conducted. An item analysis was done on two pilot exams for difficulty and discrimination of items. Both pilots were done at the University of Twente in first-year calculus courses. One pilot was in 2016 with 55 participants whom were assessed 100% on paper and also assessed on final answer (digitally). The other pilot in 2017 was a hybrid exam with 492 participants, with 2/3 digital and 1/3 written. The alignment between course goals and these exams were analysed through a policy synthesis and a content expert with the use of Bloom's Taxonomy. A focus group was conducted with lecturers to investigate their digital acceptance on testing, along with an analysis of evaluation questionnaires from the pilots. Exams were found to have a sufficient overall difficulty, discrimination and number of course goals covered. Contrary to expectations, one good digital testing item also reached the synthesis level in Bloom's Taxonomy. This item's success potentially lies in the fact that its final mark is broken down into smaller 1 and 0.5-mark increments. Findings suggest that making 50% of a final summative exam digital would be considered more acceptable among students than 2/3 digital, whilst lecturers are optimistic about the potential of digital testing in the next five years, potentially reaching up to 80% of an exam.

1. Introduction

With the introduction of technology at every level of society, opportunities within all types of assessments are increasing. Digital assessments promise being able to bring more consistency in marking, faster overall results, and instant automated personalised feedback. Ideally, all assessments seek to reflect the true ability of a student. Traditional linear, timed, unseen tests must be selective about the content that would adequately represent the learning goals of the course. The selection of test items does not only vary in terms of content per item, but also in terms of difficulty. This is done so that the test can discriminate between weak and stronger students. On the market there are digital testing products that offer summative assessment modes, as well as item question banks and formative assessment options. Unfortunately, many commercial digital testing products do not contain test items where the automated marking extends beyond correcting the final answers on an item or Multiple-Choice Questions (MCQ). One such a commercial product is MyLabsPlus, which is used at the University of Twente. As students are not able to be assessed on their argumentation and reasoning per item, this bring about concern among teachers about the extent to which these question types can bring about high-quality assessment within an undergraduate Mathematics course. In fact, some teachers might see the digitisation of assessments not as an opportunity, but as a threat. However, some argue that when testing Engineering students, there is a different curriculum that requires less argumentation and proof, and more of a knowledge on how to apply Mathematical tools.

In order to address and investigate these concerns, two pilots that were done at the University of Twente will be analysed. In 2016, the pilot test had students handed in both a paper-based and a digital version of their calculus exam. Learning from this, in 2017, a hybrid calculus exam consisting of open paper questions and closed-ended digital questions was done.

In this research, the concerns of digital testing expressed in literature and expressed by staff and students will be investigated. The focus of this research is to what extent can a calculus exam for engineering students include digital testing questions. In order to answer this, it will be investigated what constitutes a high quality closed-answer digital testing question in undergraduate Mathematics, which will be informed by the opinions of content experts and a statistical analysis of items of their reliability, validity, difficulty (p-value), discrimination and for MCQ, a distractor analysis. Curriculum alignment of digital testing questions and paper-based tests will be investigated and compared. This research will aid in defining the limitations of closed-final answer digital testing and address the acceptance of digital testing among students and lecturers. In turn, this will inform future research and pilots on digital testing in undergraduate mathematics.

2. Theoretical Framework

In order to investigate the pilots, first some background information is required. Three main sections will structure this chapter: Mathematics Education for Engineers, Reliability and Validity in Digital Assessment, and Concerns regarding Digital Testing.

Mathematics Education for Engineers

With the 21st century changing the need for skills in many subject areas, there might be a question if there really is a need for Mathematics for the Engineer in the 21st century. Current technological tools can do many complicated calculations, that once had to done by hand. Engineers no longer needs slide-rulers to do calculations but use computer programs where the Mathematics tends to be hidden (van der Wal, Bakker & Drijvers, 2017). However, just because Mathematics may be invisible, it is still vital, as one lecturer was quoted in the report by Harrison, Robinson and Lee (2005) "The mathematical ability of undergraduates is a handicap in learning mechanics" (p.20). According to van der Wal, Bakker, Drijvers (2012) even though the 21st century asks for new competencies, labelled Techno-mathematical Literacies in their paper, the need for Mathematical content knowledge has not decreased. Mathematics plays a central role in Engineering (van der Wal, Bakker & Drijvers, 2017), even though Engineers sees merely uses Mathematics as a tool. This contrasts with pure Mathematicians. According to Steen (2013) in (van der Waal, Bakker & Drijvers, 2017), in the workplace Engineers will use simple Mathematics, but need to know how to apply it in complex scenario's, whilst at Universities, usually complex Mathematics are used in simple scenarios. Kent and Noss (2002) did an investigation into the Mathematics used in the workplace, and some conclusions were the need for error detection, knowing what happens in the "black box" of your calculator, modelling and intuition. One of the interviews done in their paper says, "The aims and purposes of engineers are not those of Mathematicians" (p.5) as Engineers are not-context free and are deeply involved with modelling, design and explanation, and not mathematical structure and rigor. Some of the Techno-mathematical Literacies labelled by van der Wal Bakker and Drijvers (2017) include: interpret data literacy, which involves the analysis and interpretation of data, sense of error which involved the ability to check and verify data and technical creativity which involves creating solutions to problems. There is thus a need for good knowledge and understanding of Mathematics by Engineers, but also higher-order skills such as analysis and evaluation, but in a more context depended scenario than that of pure Mathematicians.

Many educationalists have tried to classify different educational and cognitive skills in Education. Bloom's Taxonomy (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956) is one such scheme of six categories, where Knowledge, Comprehension and Application generally classified as lower cognitive skills, and Analysis, Synthesis and Evaluation as higher cognitive skills. However, as

Radmehr and Drake (2017) warns, "some aspects of knowledge (e.g. conceptual knowledge about the Fundamental Theorem of Calculus) are more complex than certain demands of application (e.g. using Fundamental Theorem of Calculus to solve $\int_2^5 x^3 dx$)" (p.1207), thus whilst application maybe be more of a higher cognitive level than knowledge, it does not mean it is more difficult. The original Taxonomy scheme was created to help set up course goals, using those that write educational goals and those that construct tests to be aware of the verbs used, as these will reflect the educational expectation. With Mathematics having a different vocabulary of verbs, how these levels can be applied to the Calculus classroom was investigated and defined by Karaali, (2011); Shorser (1999) and Torres, Lopes, Babo, and Azevedo, (2009), showing and giving examples of how Mathematics can reach each of the cognitive levels in Bloom's Taxonomy. Radmehr and Drake (2017) explored integral calculus in depth with regards of the knowledge dimension in Bloom's revised Taxonomy. It should be noted that Karaali mentions that at the start of his research, it was difficult to think of questions that fit the higher cognitive levels and could generally only come up with questions from Knowledge to Analysis. This is the view of many, that Mathematics is an isolated rule-following one-answer only exercise (Gainsburg, 2007) making it difficult to think of higher cognitive level questions. However, there is a need for the higher order thinking skills for Engineers due to their context depended skills, and just as Karaali (2011) concludes, if one of the goals is to help students into effective thinkers, providing appropriate contexts in which they practice decision making is only reasonable. Assessing these contexts also needs to occur, as students generally learn to the test, being more influenced by that than what is taught (Gibbs & Simpson, 2005). This makes testing central to education (Evertse, 2014), and important to do well. Biggs and Tang (2011) as cited in Sangwin and Köcher (2016) state that it is important to start with the outcomes intended, and then to align teaching and assessment to these outcomes, and that all these assessments must balance this constructive alignment with what is practical, valid and reliable testing. Thus, using and evaluating the Blooms taxonomy in tests to discover the cognitive processes could be very informative, especially within digital testing, where it is thought that digital testing (MCQ) does not encourage high level cognitive processes (Airasian, 1994 & Scouller 1998 as cited in Nicol 2007).

Reliability and Validity in Digital Assessment

What makes a good assessment tool can be measured in many ways. Firstly, there is validity of the tool that is being used. Imagine wanting to measure English grammar, but then choosing an essay as your tool of measurement. Unless the marks gained for the essay is purely for grammar and not argumentation, this is not a valid means of mearing the intended outcome (Ebel & Frisbie, 2012). Other concerns in testing is the reliability of the test. Many traditional item analyses are concerned with: item difficulty, item discrimination and the distractors of MCQ's (Odukoya, Adekeye & Igbinoba, 2018; Lee, Harrison & Robinson, 2012). The next few paragraphs will explore these concepts further, in

terms of general testing, but keeping in mind final answer questions and multiple-choice questions were appropriate.

Reliability is one of the most significant properties of a set of test scores (Ebel & Frisbie, 1991). It describes how consistent or error free measurements are. If scores are highly reliable, they are accurate, reproducible, and generalizable to other testing occasions/test instruments (Ebel & Frisbie, 1991). Criterion-referenced testing, which is most of group-based testing, is not only concerned with placing students in the same order in different tests, but also that each student should achieve the same percentage-correct score across different tests (Ebel & Frisbie, 1991). Cronbach's Alpha is an acceptable measure of reliability, that can be used on both open answer and multiple-choice items (Ebel & Frisbe, 1991). What is considered an appropriate measure of reliability differs depending on what will be done with the scores. For teacher made tests, .50 is regarded as acceptable. 0.85 is needed if decisions are being made about individuals, with many published standardised tests having reliabilities between 0.85 and 0.95. If a decision is to be made about a group, 0.65 is the generally minimum accepted standard (Ebel & Frisbie, 1991).

Constructing a test, is a fine art. There is no sure way of telling what students will find difficult. It could depend on the ambiguity of the question, the reasonableness of the wrong alternatives in MCQ, or the examinees familiarity with the content (Ebel & Frisbie, 1991). Hopefully most items are also past simple recall (Myers, 1955) also causing uncertainty in how students will perform. For reliability purposes, it is argued that items should all be of the same difficulty, with around 50% of students getting it correct (Myers, 1955). However, setting a test is not only about statistics, but the test constructor is also concerned with the psychological effect it has on the test taker (Myers, 1955) and thus, for example, the exam could start off with a few easier questions. Item difficulty can be described as p-values (percentage correct). P-values describe how many students of the total, get the item correct. This value is between 0 and 1. Calculating it for non-dichotomous items involves taking the item average, divided by the maximum for the item. Whilst 0.50 may be the ideal, in reality the difficulty of items in an exam cover a great range. Many consider items with a p-value lower than 0.30 to be too difficult and should be reconsidered, and p-values above 0.7 to be too easy and should also be reconsidered. Depending on the context and purpose, these cut-off points are flexible (Odukoya, Adekeye & Igbinoba, 2018). According to Beckhoff, Larrazolo and Rosas (2000) in their testing manual in Mexico the distribution of p-values should be as follows: 5% of easy difficulty; 20% of medium-low difficulty; 50% of medium difficulty; 20% of medium-hard items; and 5% difficult items, with the median between 0.5 and 0.6.

Item discrimination can be calculated in many ways and is used to tell if an item can tell apart students of low ability from those of high ability concerning the test construct. One method for discrimination is the Item-Corrected Correlation,

to see how well an item correlates with the overall performance in an exam. Another way is the extreme group method which compares the p-value of an item of the lowest 25% to the highest 25% group. The remaining value is the discrimination index (Odukoya, Adekeye & Igbinoba, 2018). A discrimination close to 0 means that there is no different in how the lower and higher group performed, and a discrimination close to one means that everyone in the top group got it right, and nobody in the lower group got it right. This is rarely the case. Discriminations of 0.50 or higher are considered excellent (Odukoya, Adekeye & Igbinoba, 2018). According to Lee, Harrison, and Robinson (2012), a discrimination of above .40 are very good items, 0.30 to 0.39 are reasonably good but subject to improvement, 0.20 to 0.29 are marginal items usually needing improvement and below 0.19 are poor items. According to Ding, Chabay, Sherwood, and Beichner (2006) values of above .3 for the extreme value method is considered good. However, it should be investigated why an item has poor discrimination – it could be that due to a high p-value where everyone got it right. This is not always a mistake but done on purpose for a psychological boost for students and should not be removed. Another reason for a high- p-value is that it is fundamental concept that you expect everyone to get right and should be tested in the exam (Ebel & Frisbie, 1991). A low discrimination could mean a poor p-value for all and the item was too hard. This should also be investigated, whether it is it due to ambiguity or bad writing or is it genuinely a hard content question. It is thus necessary to also look at the patterns of responses, as in Multiple Choice questions the difficulty of an item also lies in the power of its distractors.

A multiple-choice question consists of a question, also known as the stem. There is one correct answer, called a key, and the rest of the options are called distractors. All of these parts together, is called an item (DiBattista & Kurzawa, 2011; Quaigrain & Arhin, 2017). Looking at the patterns responses guessing can be detected. Effectiveness of distractors to discriminate can be found through calculating the RAR values – which is the correlation between the dichotomous responses of a distractor and the responses in the exam. According to DiBattista and Kurzawa (2011) for a distractor to be good, at least 5% should choose it. If none of the distractors are chosen, the item validity could be in danger. Perhaps the item is badly written, and just by looking at the possible responses, students could guess the correct response. Then students are no longer getting a score for what the item is testing, compromising the validity of the option (Ebel & Frisbe, 1991).

Validity has been mentioned a few times as being important, and whilst it may be generally understood, the term is sometimes misunderstood or confused with reliability (Ebel & Frisbie, 1991). Thus, a definition from Ebel and Fribie, 1991: "the term *validity*, when applied to a set of test scores, refers to the consistency (accuracy) with which the scores measure a particular cognitive ability of interest." (p. 100) There are two aspects to validity, what is measured, and how consistently it is being measured, making reliability a necessary ingredient of validity. Some

analysis of questions can also provide insight into validity. Some examples of questions that have bad validity is an essay to measure grammar, and clues in MCQ's. Another concern is when the goal of the test is to measure higher order thinking, but only knowledge is being asked (or visa versa). Perhaps a mathematical concept wants to be measured, but an item needs a high level of reading and vocabulary. The last example from Ebel and Fribie (1991) is one where the instructions in an exam is to answer "True" and "False" questions using "+" or "-" but a student uses "T" and "F". If this gets marked wrong, the item is no longer measuring their mathematical ability, and it should be considered what the item is then really measuring. No score is perfectly valid or invalid, but measures can be taken to make sure we are truly measuring the intended "cognitive ability of interest".

Concerns Regarding Digital Testing

Assessments are used to make important decisions about the future of individuals, and thus it is crucial that items, whether digital or not, be handled correctly during development, administration, scoring, grading and interpretation (Odukoya, Adekeye & Igbinoba, 2018). This is true about summative assessment, which is done at the end of a course to see if a student has met a certain standard. Formative assessment is done during the course, generally as feedback to both staff and students. Digital testing is still mostly used in a formative setting (Evertse, 2014). The high stakes nature of summative testing combined with the concerns regarding the validity of using multiple choice and short answer questions – especially in the case for Mathematics – has caused this to stay so.

The advantages of multiple-choice items are numerous, but so are the challenges. As described by Odukoya, Adekeye and Igbinoba (2018), MCQ's are objective, which increases reliability (Ebel & Frisbie, 1991), and they quick to score and analyse. It is the most logical choice when assessing large groups of students and allows for a greater coverage per content in a test (Odukoya, Adekeye and Igbinoba, 2018; Chalies, Houston and Stirling, 2004; Quaigrain & Arhin, 2017). However, the downside is that the development is technical and time consuming. As Odukoya, Adekeye and Igbinoba (2018) also mentioned the challenges of writing good MCQ's as: "ambiguous prompts, poor distractors, multiple answers when question demands only one correct answer, controversial answers, give-away keys, higher probability of testees guessing correctly to mention but few of the challenges" (pp. 983-984). Concerns of validity include students eliminating options rather than working them out to the full (Nicol, 2007; DiBattista and Kurzawa, 2011). Another validity concern in Mathematics is that students can sometimes also "reverse engineer" distractors to get back to the answer, thus, you are not testing the intended skill of the item (Azevedo, Oliveira, & Beites, 2017). Setting good distractors is difficult (DiBattisa & Kurzawa, 2011). In the development of multiple-choice items, requiting the relevant subject experts is a crucial step in the item validity of a question. The writer needs a good knowledge of the content being assessed, an understanding of the objectives of what is being

assessed (Vyas & Supe, 2008). However, this will not guarantee validity, and trail testing of items is required, along with statistical analyses (Odukoya, Adekeye & Igbinoba, 2018). Due to all these complexities, it is also rumoured that MCQ benefits the average student, and disadvantages the stronger students (Sangwin & Köcher, 2016). There are also many Multiple-Choice taxonomies that can be followed for writing reliable and valid items (Torres et al., 2009; Haladyna, Downing, & Rodriguez, 2002; Burton, Sudweeks, Merrill & wood, 1991) however these are not specific to Mathematics, in addition to there being widespread ignorance of such frameworks (DiBattista & Kurzawa, 2011).

There are concerns about to what level of difficulty and validity digital testing questions can offer to Mathematics. Mathematics is traditionally assessed on paper, and marks are given for the method. In order to make a question "fair" in digital testing which can only assess "final answers" and not method, the questions do not require long complex calculations (as done in Chalies, Houston & Stirling, 2004). They should also test one part of a set task - bringing about the concerns heighted by Lawson (2001) in Paterson (2002) such as digital testing questions only test lower cognitive skills (Kastner & Stangl, 2011), they provide more information in the question to the testee and force a method on the user. In a report by Everste, (2014) there are doubts whether or not digital testing can test higher order thinking. DiBattista and Kurzawa (2011) and Quaigrain and Arhin (2017) state that it is possible to write a Multiple-Choice question that tests higher cognitive skills, but this requires a lot of skill from the item writer. Hoffmann (1962) is quoted in Sangwin and Köcher (2016) to have said that Multiple Choice "favour the nimble-witted, quick-reading candidates who form fast superficial judgements" and "penalize the student who has depth, subtlety and critical acumen" and many continue to have this critical look on the item type, as Torres et al (2009) notes that many teachers have the idea that multiple choice "can measure only memory, and does not give students the necessary freedom of response to measure more complex intellectual abilities"

Computer Algebra Systems can evaluate final answer questions in relation to a string of accepted answers. Final Answer marking is very well established and can assess anything from matrices to equations (Sangwin & Köcher, 2016). However, other concerns about the digital testing come from students about losing all their marks when making a small mistake, wrong input or wrong rounding (Chalis, Houston & Stirling, 2004), and with that thus the lack of partial credit for items (Chalis, Houston & Stirling, 2004; Naismith & Sangwin, 2004). Students cheat more easily (Azevedo, Oliveira & Beites, 2017) and in reaction to cheating Chalis, Houston and Stirling, (2004) suggests that parameterisation is crutial. However, as Impara and Foster states (2006), strategies to reduce cheating in digital exams "What makes for good security does not always make for good psychometrics" (p.95)

3. Research Questions and Hypotheses

Digital testing is domain specific (Kastner & Stangl, 2011). Only research conducted in undergraduate Mathematics can contribute to answering questions that come with the complex task of mass testing at University level. Assessment forms an important part of education, and as put succinctly by Ridgeway, McCusker and Pead (2004) in Sangwin and Köcher (2016):

> The issue for e-assessment is not if it will happen, but rather, what, when and how it will happen. E-assessment is a stimulus for rethinking the whole curriculum, as well as all current assessment systems.

Considering the concerns and uncertainty concerning summative digital exams for Mathematics, our research question is:

> *To what extent can digital testing be included in first year calculus summative exams, for Engineering students?*

Sub-questions

In order to answer the main research question, the following questions are investigated:

1. Which differences are there in digital or written questions in meeting course goals at different cognitive levels?

2. To what extent can digital testing questions create the expected distribution of students according to their mathematical ability?

3. How is overall and item-wise discrimination effected by digital testing?

4. What is the current state of acceptance of digital testing calculus amongst staff and first year engineering students?

Hypotheses

Hypothesis 1: It is expected that digital testing will be able to meet a variety of course goals – however it is expected that it will only cover memory recall (Knowledge) and basic procedures (Understanding) within the levels of Bloom's Taxonomy.

Hypothesis 2: It is expected that digital testing question will be of a lower p-value, causing a distribution that might represent a normal distribution curve, but shifted to the left of the written distribution curve. The shape of the curve will have less of a standard deviation, due to the digital testing questions having less of a variety of difficulty.

Hypothesis 3: Discrimination of items has to do with the sorting of groups. It is expected that the mid-achieving students will have the greatest disadvantage from digital testing questions, having the greatest difference in marks. On an item level, distractors should be chosen by weak students and avoided by strong students. It is expected that distractors of common misunderstandings that students make will be effective in doing this.

Hypothesis 4: It is expected that staff and students might be open to digital testing questions, but only for the basic skills covered in an exam.

4. Method

4.1. Context

The University of Twente is set between two suburban cities, Enschede and Hengelo, in the province Overijssel in the Netherlands. The University of Twente (UT) was started in 1961 and has the motto *High Tech, Human Touch.* The University works with other technical Universities through the 4TU federation: Wageningen, Delft and Eindhoven.

In 2015 a project group formed at the UT, called "*Project Digital Testing".* The group started with Steffen Posthuma as project leader, program director Jan Willem Polderman, as the client, Jan van der Veen as the chairperson of the 4TU Centre for Engineering Education, as the financier and supports the project with expertise. Karen Slotman as an expertise in testing, from the Centre of Expertise in Learning and Teaching (CELT) and Harry Aarts as expert in Maths education. As additional support, the project team also has as per consultation: Bernard Veldkamp as consultant in methods and techniques in statistical analysis, as an expert in adaptive testing. Stephan van Gils and Gerard Jeurnink as lecturers of Mathematics X[1] (Calculus), Brigit Geveling from Applied Mathematics, as well as representatives of Electrical Engineering (EE), such as the Examination Committee.

In 2017 Anton Stoorvogel and the researcher joined the research group. Anton Stoorvogel is a Mathematics professor, supporting the research group in their third pilot, focusing on expanding their pilots from Calculus to Linear Algebra.

This Project endeavours to run pilots regarding summative digital testing in first year mathematics courses for engineering students. The program MyLabsPlus from Pearson has been up to this point been used for formative testing. The aim of these pilots is with regards to summative testing: "To what extent can Maths X be digitally tested with MyLabsPlus". They had two criteria for quality questions, which is validity and reliability. By the end of the 2017 academic year, the project team had run three pilots, the first two in the subject area of first year Calculus and then one in first year Linear Algebra, with each pilot building on knowledge gained from the previous pilot. The project has, in the academic year of 2017/2018, access to 150 Chromebooks that can be used for secure digital testing.

This research project makes use of the pre-existing data sets from the first two pilots run by the project team in the subject area of first year Calculus for Engineering students. The first pilot made use of exam questions from an item bank and the second pilot had questions made by the project group, based on feedback from the first pilot.

[1] Course name has been changed to Mathematics X for privacy reasons.

4.2. Respondents

Respondent during the 2016 pilot.

Sampling procedure used was in-tact sampling as it was done at University where classes could not be split. Three practical reasons determined Maths X being chosen for sampling was done for the pilot test, which includes the good class size for statistical purposes within the EE class, as well as Maths X has many textbooks available to generate and search questions from - which students have not seen yet, and permission from the examining committee to run a pilot with the EE group within the Maths X line. The participants in the 2016 pilot were thus first year EE students. The nationality mean age and gender of the participants is not available due to the privacy laws at the University of Twente. As all participants were in their first year of University, it can be assumed that they are of the average age of 19. It can also be assumed that most of the students were of Dutch nationality.

Respondents during the 2017 pilot.

The 2017 pilot built on conclusions made from the 2016 pilot, and thus the same first year-calculus course, Maths X, was chosen for the pilot. The 2017 pilot was a much larger pilot with 492 participants from various studies. The studies included in the pilot were: Electrical Engineering, Mechanical Engineering, BioMedical Engineering, Software Technology, Advanced Technology, Civil Engineering, and Industrial Engineering and Management. In-tact sampling was used where one group (n = 52) did the pilot on Chromebooks, and the rest of the students (n = 440) wrote the exam on equivalent paper-based versions. The nationality, mean age, and gender of the participants is not available due to the privacy laws at the University of Twente. It can be assumed that the majority of the students are Dutch. As all participants were in their first year of University, it can be assumed that they are of the average age of 19. Both groups were given a voluntary questionnaire at the end of the study. The EE group were given additional questions regarding the digital aspect of the exam.

Content Experts Respondents for Interviews.

Content experts regarding teaching, calculus and digital testing item construction were consulted at various stages of the thesis. Content experts were chosen either due to their involvement in the digital project group, through recommendations made through the project group or through a reading group that the lecturer attended.

Respondents for the Focus Group.

This focus group was conducted with six lecturers from the University of Twente. The recommended list of respondents was a list of lecturers that were involved in first year calculus courses. The researcher requested the list from the project group. Thirteen lecturers were sent an email (appendix 5), which explained the purpose of the focus group, the time, date and that free lunch will be provided, and how the focus group will be recorded and kept confidential. If lecturers could

not come, but wanted to give an opinion, they were invited to send an email. One lecturer made use of this opportunity and provided his opinion and years of experience. This response is seen in appendix 7. The final group of participants for the focus group were six lecturers. One lecturer, however, is already a part of the focus group. The experience of lecturers in teaching ranges from 2.5 to 38 years (17.4 years average). Not all experience is at the University of Twente, but also at other Universities. Only one of these lecturers has never had any of their courses tested digitally. Most lecturers had some experience with some of their courses being tested digitally, with Multiple Choice being mentioned the most. Experience ranged from making items in Maple TA, having half a course tested with Multiple Choice for 6 years and organising the digital platform for a university. Two others besides the researcher were also present that mainly posed questions: an educational advisor for the mathematics faculty and an associate professor from ELAN (Department of Teacher Development) that is and has been involved with digital testing projects and assisting lecturers that want to adopt digital testing in their courses.

4.3. Instruments

Evaluation Questions as Instrument.

Evaluation questions for both the 2016 and 2017 pilot, were brainstormed written by the project group. An example of a question from the 2016 evaluation is "I believe that a digital Math exam with MyLabsPlus using the Respondus Lock Down Browser is a good way to test my knowledge and skills." This questions changed slightly in the 2017 evaluation questions due to the change in the format of the pilot, to: "I believe a hybrid Math exam with both short answer questions (e.g. multiple choice) and open questions with written solutions (incl. calculations) is a good way to test my knowledge and skills". All questions asked can be found in appendix 2 and appendix 4, for the 2016 and 2017 evaluation questions, respectively.

Exam Questions as Instrument.

The calculus questions during the 2016 pilot were taken directly from the item-bank in the MyLabsPlus program. Students answered these questions on paper, as well as on the Chromebooks. These questions were thus existing questions from Pearson. The pilot consisted of nine questions – consisting of a total 14 items of which 12 are final answer and two are multiple choice. The exam questions can be seen in appendix 1.

The calculus questions that were used during the pilot in 2017 were designed by content experts at the University of Twente. The pilot consisted of two paper-based questions, six multiple choice questions and five final answer questions. The digital testing component consisted of two-thirds of the marks, and the written component one third of the marks. The exam questions can be seen in appendix 3.

Focus Group Questions as Instrument.

The focus group was organised and developed by the researcher. Questions of interest were brainstormed together with the project team. Four main questions were identified beforehand that could be asked during the focus group: "What is your first impression regarding the advantages of these question types for Mathematics, as in the 2017 pilot?", "Would you use these question types in an exam that you were setting? If not, why not?", "What possibilities/question types would you like to see in digital testing?", and "Hypothetically speaking: Say that digital testing becomes the norm at the University, what kind of support would you as a lecturer would like to receive?" The structure and brainstorming of questions can be found in appendix 6.

Curriculum goals as Instrument.

The curriculum goals of the first-year calculus course, Math X, is presented in different documents. "Educational Targets" was chosen for this research, describing the main educational goals that should be reached in the course. These educational goals can be seen in appendix 8. Other documents that are not used are "Course description" which describes the position of the course within the other mathematics course before and after the course and in the "Schedule of topics" describes how the course is structured according to chapters in a textbook.

4.4. Research Design

An ex-post facto design was adopted for this study, as secondary data was collected and analysed. Research done for this study will be mixed methods, consisting of quantitative statistical data (test scores and Likert scale answers) and qualitative data from open answer questionnaires from the two pilots conducted in 2016 and 2017, and a focus group that was conducted by the researcher with lecturers from the University of Twente.

In the 2016 pilot, students could decide to not participate in the pilot. 65 wrote the exam, but only 56 consented in their data being used for a pilot. All participating students except for one filled in all the evaluation questions.

In the 2017 pilot, the evaluation questionnaire was not fully filled in by all attending, resulting in a variety of responses for each question in the questionnaire, with a minimum of 330 (66.8%) and a maximum of 373 (75.5%). From the 52 participating EE students, 44 (85%) filled-in the evaluation questionnaire.

4.5. Procedure

Procedure during the 2016 pilot.

The digital testing pilot occurred in week 24, on 10 June 2016. Students were informed beforehand if they had objections to participating in the pilot, they could email the project team. The arrangements for the pilot were as follows: During a normal Maths X exam, pilot students took their seats in the middle of the room with two-persons tables. Students were instructed to first do the two-hour handwritten exam and thereafter enter their answers in MyLabsPlus. 15 minutes

before the end of the exam, students were given a sign to start entering their answers in MyLabsPlus. Students would open the test on Blackboard, which resulted in their laptop being locked-down - meaning that they could not access any other part of their laptop or internet until the end of the exam, a measure in preventing cheating. A gift coupon of 10 euro was awarded to all the pilot students whom entered their answers in the MyLabsPlus program and answered the evaluation questions. If the laptops of the students did not work, there were back-up UT laptops available. Alternatively, there were paper-copies of the digital testing available if working online would fail altogether. Student Assistants were in the room to check that the correct summative test was started up, and not one of the diagnostics tests used earlier in the course. Each student assistant would survey a block of students, in addition to an invigilator at the front for questions.

Data was collected through the MyLabsPlus programme, as well as through written exams. Written exams were marked as normal - through the use of an answer scheme by experienced lecturers. The MyLabsPlus exams were graded automatically through an electronic grading scheme - only based on the final answers entered or the multiple-choice option selected. As it was a pilot for digital testing, students were graded on their written exam, and not the equivalent digital exam. The final data of how many marks each student got for the paper, and the digital exam, and evaluation questions was collected and it was entered into an Excel spreadsheet by a student assistant. Before being imported into SPSS for analysis, The researcher recoded Questions 2 to 11 to be on the same scale of 1 = totally disagree and 5 = totally agree, as in the 2017 pilot.

Procedure during the 2017 pilot.

On the 15th of May 2017, students were invited to participate in a diagnostic test to get used to the new format of the module exam on the 16th of June, which would consist of open questions, multiple choice questions and final answer questions. During the diagnostic test, Electrical Engineering students were provided with a Chromebook, just as they would be during the final exam. The diagnostic test was not compulsory, but the presence of students was highly recommended.

On the 16th of June, 492 students participated in the pilot which was an exam for Maths X. The exam consisted of 36 marks, of which 33% marks were open, written, questions, 42% were multiple choice questions and 25% were final answer questions. Of the 492, 52 Electrical Engineering (EE) students wrote the multiple choice and final answer questions on a Chromebook. Access to all other software on the Chromebook and internet were blocked. All other students also wrote the multiple choice and final answer questions, but on an equivalent paper-based version. Evaluation questions for both groups were optional.

The data collection of the two-thirds digital component for the EE students were done using MyLabsPlus, which were graded automatically through an electronic grading scheme - only based on the final answers entered or multiple-choice option selected. The paper-based exams were marked by experienced lecturers, using an answer scheme. Informal contact occurred between lecturers

for consistency through marking together or checking answers through WhatsApp groups. It was ensured that lecturers, despite the final answer questions and multiple choice being on paper, would mark the answers as a computer would do it, resulting in reliable data processing, making it possible to analyse the data as if it was assessed digitally on a Chromebook.

The final data of how many marks each student got for each question and the evaluation questions were collected and entered an Excel spreadsheet by a student assistant - thereafter it was imported into SPSS for statistical analysis by the researcher.

Procedure during the Focus Group.

An email (appendix 5) was sent to invite lecturers from the University of Twente to a focus group regarding digital testing. A focus group was conducted with six lecturers from the University of Twente, along with two other members of the digital testing group. On the day of the focus group, the room was open by 12:15 where lunch and refreshments were ready. The focus group started at 12:30, where lecturers were seated and were welcomed by the researcher. The researcher did a short introduction about the purpose of the research, handed out confidentiality forms as in appendix 6 as well as handed out the digital testing items of 2017 (appendix 3). The full plan of the focus group can be seen in appendix 6. The focus group was a relaxed semi-structured meeting where the researcher mainly posed questions and follow-up questions, but those present from the focus group also posed some questions in line with future possibilities. The focus group ended at 12:25 as some needed to leave for meetings – however the majority still stayed until 12:40 as they were engaged in conversation. In the email inviting lecturers to the focus group it stated that focus group will be recorded and that what they say will be treated confidentially, as they will only be identified through their years of teaching experience and experience in digital testing. The focus group was thus recorded using two recorders owned by the researcher: a phone and a tablet. The recordings were only accessible to the researcher and were used to make a summary. Notes were taken by two members of the project group during the meeting and sent to the researcher to aid in the reliability of the summary. A final summary was made by the researcher (appendix 7) and was sent back to all that participated and given a week to respond if they wanted to express any more opinions or disagree with something in the summary. No-one responded.

5. Analysis and Results

First how the data were analysed is described, followed by the results of the analysis. See table 1 for an overview of resources used per research question.

Table 1
Overview of Resources Used for Each Sub-Question

Research Questions	Pilot 1 and Pilot 2			CE	PS	FG
	Questions	Exam results	Eval			
RQ1_CognitiveLevels	x			x	x	
RQ2_Difficulty		x				
RQ3_Discrimination		x				
RQ4_DigitalAcceptance			x			x

Note. CE = Content Expert ; PS= Policy Synthesis ; FG = Focus Group

5.1. Data Analysis

As the 2017 pilot learned from the 2016 pilot, how data is presented is as follows within each sub-question: (1) How the 2016 pilot is analysed, (2) How the 2017 pilot was analysed and then, (3) What can be learned from comparing both pilots.

Data analysis of sub-question 1:
Which differences are there in digital or written questions in meeting course goals at different cognitive levels?

This sub-question was answered with the help of a content expert that rated the questions of both pilots, with the help of a coding scheme from literature, and Educational Target documentation.

Data analysis for sub-question 1 using the 2016 and 2017 pilots.

A content expert did all the questions fully as a student would, and then coded each question with the cognitive levels of Bloom's Taxonomy, and with one or more Educational Targets from Mathematics X. The cognitive level of each question was checked using a Bloom's Taxonomy devised for Mathematics by Shorser (1999) and Torres et al. (2009), This taxonomy provided the content expert with a definition, an example and keywords. The full taxonomy used for the coding can be seen in appendix 9. The original Educational Targets were coded from 1.1 to 1.12 and 2.1 to 2.5 for ease of coding, and the numbers depended on whether the target concerns working with partial derivatives and applications or double and triple integral over bounded regions, respectively. The educational targets were otherwise unaltered. This information is organised in a table in terms of the levels in Bloom's Taxonomy and included whether these questions are asked using open written, multiple choice or final answer questions. Course Goals could not be labelled with Bloom's Taxonomy as they were not written with Bloom's Taxonomy in mind, mostly having the verb "Apply". This research was thus discarded.

Data analysis for sub-question 1 by comparing the 2016 and 2017 pilot.

The results from the exams were compared, to see if there are any differences between the digital exams that were made using an item bank (2016), or those that were made by content experts at the University (2017). It was counted and recorded how many course goals are covered in the exams.

Data analysis of sub-question 2:
To what extent can digital testing questions create the expected distribution of students according to their mathematical ability?

This sub-question was answered through the analysis of the results of the 2016 and 2017 pilot exams, calculating percentage correct values (p-values). P-values are divided into five categories as seen in Table 2. In the context of this thesis, where students are also given projects and their marks are not only based on a written exam, a very easy item is considered to be above 0.8. See Table 2 for interpretation of other categories into p-values. 90% of the exam should have values between .30 and .80 for optimal discrimination, not including the 5% very easy to give students confidence and %5 very hard to tell the top students apart from average students.

Table 2
Ideal P-Value Distribution in an Exam

P-values	Category	% of Exam
$P \leq .30$	Very difficult	5%
$.3 < P \leq .45$	Mildly difficult	20%
$.45 < P \leq .65$	Average	50%
$.65 < P \leq .80$	Mildly easy	20%
$P > .80$	Very easy	5%

The 2016 pilot was in the academic year of 2015-2016, and the pilot of 2017 in the year of academic year of 2016 – 2017. In order to get an idea of the difficulty of the different exams, the pass rate of the two academic years and these exams were compared. 2013 – 2014 had 407 participants, with 75% passing. 2014 – 2015 had 500 participants, with 71% passing. 2015 – 2016 had 457 participants, with 88% passing. 2016 – 2017 had 494 participants with 77% passing

Data analysis for sub-question 2 using the 2016 pilot.

In the pilot from 2016, it was firstly checked if there is a significant difference between the mean of paper-based results against the equivalent questions tested digitally using a paired sample t-test. The overall scores of the questions assessed on paper or digitally were P-value was calculated by the Mean divided by the maximum possible mark. Reliability of the exam as a whole will be conducted using Cronbach's Alpha and each item was analysed for the Cronbach

Value if the item were to be deleted. Cronbach Alpha for digital exam is .58 and Cronbach's Alpha for the written exam is .68

Data analysis for sub-question 2 using the 2017 pilot.

In the 2017 pilot, the MCQ, final answer- and open questions had the difficulty of the item evaluated using p-values, The average p-value for each type of item was compared using a paired-sample t-test to discover if there are significant differences between the means of the different item types. The different levels of p-values per item was be cross-tabulated with the three question types for analysis.

Reliability of the exam investigated using Cronbach's Alpha and each item was analysed for the Cronbach Value if the item were to be deleted. In the 2017 exam all missing values were filled in with a zero to ensure accurate processing by the software.

Data analysis of sub-question 3:
How is overall and item-wise discrimination effected by digital testing?

This sub-question will be answered through both pilots. The 2016 pilot will give insight into the difference in discrimination between paper and digital testing items, as well as what happens to overall groups in a test in a digital testing exam. The 2017 pilot focusses more on the distractors of multiple-choice items, and how well these discriminate between students of different abilities. In both pilots, discrimination is also measured through the calculating the corrected item-total correlation (CITC) and through the extreme group method which measures the item-criterion correlation, which subtracts the p-values of the bottom 25% from the top 25%, to measure the internal consistency of each item. Items with above .40 are very good items, 0.30 to 0.39 are reasonably good but subject to improvement, 0.20 to 0.29 are marginal items usually needing improvement and below 0.19 are poor items. Similarly, the RAR values of the distractors in the 2017 exam will also be analysed. In addition, if a distractor is chosen more than 5%, it is considered good.

Data analysis for sub-question 3 using the 2016 pilot.

For discrimination a Corrected Item-Total Correlation was performed per item, as well as a taking the difference between the p-value of the top 25% performance group from the bottom 25% group. That is, the low performance group (based on the written exam) is the bottom 25% scoring 58.7% or lower (n = 14) and the high performance group is the top 25% scoring 83.3% or higher (n = 16). For the groups, a scatterplot was created for analysis. The final mark of students according to the paper-based version were plotted on the x axis, and the marks of students according to the digital exam on the y axis. A few reference lines follow: The y=x line was plotted as scores near this line indicate no difference between the digital and paper exams. Eight lines parallel to this line were plotted at 0.5 intervals, four above and four below, to show how far students deviate from the y=x line. Thus these parallel lines have equations $y = x + 0.5$; $y = x + 1$; $y = x + 1.5$; $y = x + 2$ and $y = x - 0.5$; $y = x - 1$; $y = x - 1.5$; $y = x - 2$. Two

more reference lines are present, showing the cut off points for pass: y = 5.5 and x = 5.5. Interpretation of the graph will occur using these reference lines to see if there are specific groups that deviate further away from the y=x line than others. It will be investigated how many students fail using a digital test in comparison to a written test, and visa-versa. This will be done by looking at the shape of the graph and by counting.

In addition to the scatterplot, t-tests were conducted to discover if the level of academic achievement has an influence in whether students do better in digital or paper-based exams. The results from the 2016 pilot was split into three groups of low, medium and high achievement according to the written exam. Each group contained 33% of the group. The mean P-values for each item for each group for both the written and digital components was be calculated. For each item, within each academic group, a paired sample t-test was done to compare the significance between the mean of the written and digital tests. This is used to see if there are specific items that cause the main differences between digital testing and paper-based questions, and within any academic group.

Data analysis for sub-question 3 using the 2017 pilot.

For discrimination a Corrected Item-Total Correlation was performed per item, as well as taking the difference between the p-value of the top 25% performance group from the bottom 25% group. That is, the two group discrimination was calculated through subtracting the p-value that the bottom 25% (academically) had for a question from the top 25%. The first quartile was cut off at 50% for the exam (n = 126), the second quartile at 63.89% and the top quartile was above 76.39% (n = 125).

The percentage of how many students chose a distractor will be analysed for every multiple-choice item, in order to determine effective distractors. In addition, the RAR values for each of the distractors will be analysed. This information, along with the information from the two previous sub-questions will be used to see what the difference between high quality items and their distractors is, versus low-quality items and their respective distractors.

Data analysis of sub-question 4:
What is the current state of acceptance of digital testing calculus amongst first year engineering students and staff?

This sub-question was answered using the results from the both the Likert Scale answers by students as well as the open answer questions from both Pilots in 2016 and 2017, as well as a focus group of staff members. A construct of "Digital Acceptance" was created in both pilots as a qualitative measure of acceptance. A result of above 3.5 on a 5-point scale is considered to be good, a result between 2.5 and 3.5. to be reasonable and below 2.5 to be worrisome. In order for the results to be comparable between pilots, the 2016 evaluation questions were recoded, so that 1 = totally disagree and 5 = totally agree, just as in the 2017 pilot. Cronbach's Alpha was used to calculate the reliability of the items chosen for the "Digital Acceptance" construct. The open evaluations questions were coded by two raters in order to count and categorise comments made by students

regarding digital exams in both pilots. The focus group was recorded and summarised by the researchers and was checked by those that participated.

Data analysis for sub-question 4 using the 2016 pilot.

In order to construct a digital testing acceptance construct amongst students, the evaluation questions were inspected. Question 8 asks if digital testing is a good way of testing and question 9 asking if digital testing would be a fair way of testing. It was decided that these two questions were indicative of acceptance of digital testing. An explorative factor analysis was used to discover other variables that aligned with these. A factor loading higher than .40 would be considered for a factor. Questions 4, 5, 8 and 9 were considered for the factor "Digital Acceptance" as they had a factor loading of .49 .91 .45 and .48 respectively for a single factor. Question 4 asks students if it would be fair to have their digital answers as their final mark, and not their written answers, and Question 5 asks if the student would find it advantageous if only their final answers were assessed, and not their calculations. A Pearson Correlation was conducted on questions 4, 5, 8 and 9 to confirm that these questions make one factor, as in the factor analysis three of these were above the .40 cut-off mark, but still below .50. Question 4 and 5 correlated with .44 with a significance of 0.01, and question 8 and 9 correlated with .62 with a significance of 0.01. All correlations between the four questions were at least significant at the 0.05 level with the lowest correlation of .31. The table can be seen in appendix 10. The Cronbach's Alpha of the four questions was .72, which is highly acceptable, which is good. However, it was decided to rather remove Question 5, shortened to "DigiOnlyAdvantage", from the construct as the question might correlate well, however neither a high or low score objectively tells much about digital acceptance of a student. For the other three questions making the construct, a high value means a high acceptance towards digital testing, that is, believing that digital testing is good and fair way of testing. Since question 5 has the word "advantage" it is testing if the students believe that they would score more marks in a digital exam than on a written exam. This is not the same as believing that digital testing is fair. Ideally, it is favourable that this score is neutral around a score of 3, showing that the student believes that the written and digital versions of the exam are comparable. It was checked to see how much a difference it would make removing this from the digital acceptance construct. With four variables, which included question 5, the digital acceptance construct "DigiAcceptance" only had 0.04 more in the mean than if three variables were to be used, and the standard deviation would only be 0.04 less. The Cronbach's alpha with three variables it was .70, which is still very acceptable. It was thus decided to discard question 5 from the digital acceptance construct, and construct it using evaluation questions 4, 8 and 9.

Each evaluation question in the pilot was analysed separately for mean and standard deviation for an overview of the questions. Question 1 was excluded from this analysis, due to the different scale, and only concerning how fast students think they can answer questions in MyLabsPlus (MLP).

Data analysis for sub-question 4 using the 2017 pilot.

In order to create a variable for digital testing acceptance among the students in the 2017 exam, the questions were inspected. Questions 4, 5 and 6 were re-coded first, so that a high score among all items would mean a high acceptance towards digital testing. An exploratory factor analysis was done with the first seven questions that were presented to all 492 students. The factor analysis resulted in two factors, and the one factor appeared to measure "digital testing acceptance" with Questions 1, 2 5 and 6 being having a factor loading of .83, .70 .84 and .59 respectively. Questions 1 asked an opinion about if digital testing is good, questions 2 about if it is fair, question 5 allowed students to indicate the percentage that they preferred being digital in an exam and question 6 asked students if they believed they would have scored better in an traditional exam, rather than the hybrid exam they wrote. A reliability analysis showed a Cronbach's Alpha of 0.82, which is good. A table showing the full factor analysis and an additional factor analysis, along with a detailed Cronbach's Alpha, is found in appendix 10.

Some attempts were done with the 2017 evaluation questions among EE students only, and then including the "Q15_PreferDigitalExam" question but due to the dataset being below 50 participants, a factor analysis cannot be run. This way of analysis was therefore discarded.

Each evaluation question in the pilot was analysed separately for mean and standard deviation for an overview of the questions. Question 10 was excluded due to a different scale as it asked students to select any hardware issues during the exam. This falls outside the scope of the study. Evaluation question 5 in the 2017 pilot asks students to indicate their preference in percentages for how much should of the exam should be digital. This evaluation question was analysed separately as a direct answer to the main research question. This question has the wording "The ratio between short answer questions (2/3rd of all points) and open questions with written solutions incl. calculations (1/3rd of all points) in this exam is right.". However, in the 2017 pilot this was in a different position for the EE students and the paper group, as well as the paper group having a five-point Likert scale, whilst those on the EE students answered it "Yes" or "No". The researcher decided to therefore combine them, converting "Yes" and "No" to "Agree" and "Disagree". In the paper-based questionnaire, those that filled in that they agree that the ratio was right, could still give a preference for the ratio they wanted, whilst this is not true in the electronic questionnaire filled in by the Electric Engineering Students. Due to this question being directly relevant to the answering of the sub-question and due to this difference, a detailed table was done with this question.

Data analysis for sub-question 4 by comparing the 2016 and 2017 pilot.

The digital testing acceptance between the two pilots was be compared to see if the changes made from the 2016 exam to the 2017 exam had any influence. To see if the difference is significant, an independent sample t-test was conducted. It was also investigated if there is a difference in the mean of digital acceptance

amongst low, mid and high achievers according to their overall exam result in the pilot, where a one-way Analysis of Variance (ANOVA) was conducted.

In addition to the Quantitative answers, both pilots also have open answer questions. The responses were read multiple times by the researcher to identify keywords, in order to make a code frame. The comments by the students were coded in whole and to keep the meaning conveyed intact. For reliability the responses were coded by an ex-student that has achieved a master's degree in Educational Science and Technology. An inter-rater reliability of 0.87 was achieved. The process of the making of the code frame, and the calculations for the inter-rater reliability can be seen in appendix 11.

A focus group with staff from the University of Twente was conducted, recorded and summarised. Findings from this is presented in the results section to indicate the current state of digital testing acceptance is with staff at the University of Twente. One lecturer could not make it to the focus group but decided to send his comments per mail. These are also included in the results section.

5.2. Results

Results sub-question 1:
Which differences are there in digital or written questions in meeting course goals at different cognitive levels?

Whilst 2016 was analysed first, and then 2017, it is more effective for this sub-question to first present comparisons between the two pilots, and then a detailed table for each sub-question.

Results for sub-question 1 by comparing 2016 and 2017 pilot.

Table 3 contains the Educational targets from Mathematics X, and which of these goals have been met in the pilot from 2016 and 2017. These have been checked independently by a content expert. The exam from 2016 meets 11 of the 17 course goals (65%), and so does the 2017 exam. However, the 2017 exam contained two written questions which met two course goals that was not covered by the digital testing questions, making the digital questions that cover 24 out of 36 marks (67%) cover 9 of the 17 course goals (53%).

Table 3
Educational Targets for Mathematics X in the 2016 and 2017 Pilot

Code	*Educational Target*	*2016*	*2017*
1.1.	Apply the parametrization of a curve and the tangent vector	✓	✓
1.2.	Apply the chain rule (in several forms)	✓	✓
1.3.	Calculate a directional derivative, and apply its properties	✓	
1.4.	Calculate the gradient (vector)	✓	✓
1.5.	Apply the relations between gradient and level sets		✓
1.6.	Calculate the tangent plane and normal line		
1.7.	Apply a linearization (standard linear approximation)		✓
1.8.	Estimate a change using differentials		
1.9.	Calculate Taylor polynomials (first and second order, two variables)		
1.10.	Apply the first and second derivative tests	✓	✓
1.11.	Calculate the absolute extreme values on closed bounded regions	✓	
1.12.	Apply the method of Lagrange multipliers	✓	✓[a]
2.1.	Sketch the region and find the limits of integration	✓	✓
2.2.	Calculate an iterated integral (by changing the order of integration)	✓	✓
2.3.	Define area, volume, mass or the average value as an integral	✓	✓[a]
2.4.	Apply polar, cylindrical or spherical coordinate substitutions, or a given transformation	✓	✓
2.5.	Calculate centroid, (center of) mass and first moments		
Total goals met		11/17	11/17

Note. [a]Only met in the written component of the exam

Table 4 contains how each level of the Bloom's taxonomy, that was present, was met in the 2016 and 2017 exam. Each exam is composed out of 36 marks; however, percentages are used for comparison purposes. The full classification scheme for what is meant by each level of Bloom's Taxonomy can be found in appendix 9.

Table 4
Bloom's Taxonomy Across Final Answer (F), Multiple Choice (M) and Written Answer (W)

Bloom's	2016			2017			
	F	M	Total	F	M	W	Total
Synthesis	6%		6%				
Analysis	17%		17%			17%	17%
Application	11%		11%		22%		22%
Comprehension	42%	3%	44%	17%			17%
Knowledge	17%	6%	22%	8%	19%	17%	44%

Results for sub-question 1 from the 2016 pilot.

Table 5 contains a detailed table of how each of the Bloom's Taxonomy level is met in the 2016 exam, with each question, the item type, the maximum marks, the concept covered and the educational targets. IT should be noted that this exam has question 9 at the Synthesis level and question 7 ant the Analysis level.

Table 5
Detailed 2016 Pilot Questions According to Blooms Taxonomy

Blooms	Question	Concept	Goals	Mode	Mark Breakdown
Synthesis	9	New Integrand and Limits	2.4.; 2.1.	F	4 of 0.25 and 1 of 1
Analysis	7	Volume cylinder	2.3.; 2.4.	F	6
Application	6	Vertical/horizontal Integration Limits	2.1.	F	8 of 0.5
Comprehension	3	Local maxima and minima	1.4.; 1.10.	F	3 of 2
Comprehension	4	Max and Min with constraint	1.12.; 1.4.	F	2 of 2.5
Comprehension	5	Area of Integration	2.1.	M	1
Knowledge	1	Vectors	1.1	F	9 of 0.33 and 1 of 1
Knowledge	2	Directional Derivative of function at point	1.3.; 1.4.	F	3
Knowledge	8a	Jacobian	N/A (2.4)	F	3
Knowledge	8b	Sketch under transformation	2.4.; 2.1.	M	2

Results for sub-question 1 from the 2017 pilot.

A similar detailed table for 2017 is presented as table 6 as some final answer questions also reached the application level of Blooms Taxonomy. These were questions 7 and 9. The table also shows how the written question 10 is at level of analysis, whilst the written question 6 is at the level of knowledge. Multiple choice questions in the application phase are 2, 4 ,7 11b.

Table 6

Detailed 2017 Pilot Questions According to Blooms Taxonomy

Blooms	*Question*	*Concept*	*Goals*	*Mode*	*Mark Breakdown*
Analysis	10	Volume Cylinder (Triple Integral)	2.1.; 2.2; 2.3 ; 2.4	W	6
Application	2	Chain rule	1.2.	M	3
Application	4	Equation Tangent Line	1.5.	M	2
Application	7	Region of Integration	2.1	M	1
Application	11b	Image Integration under Transformation	2.1; 2.4.	M	2
Comprehension	9	Change Order Integration Limits	2.1.	F	1 of 0.5 and 6 of 0.25
Comprehension	12	New Integrand under Transformation	2.1.; 2.4.	F	4 of 0.5 and 1 of 1
Knowledge	1	Vectors	;1.2.	F	4 of 0.5
Knowledge	3	Gradient Vector	1.4; 1.5.	M	2
Knowledge	5	Critical Points	1.10	M	4 of 0.75
Knowledge	6	Optimisation/Lagrange	1.12.; 1.4.	W	6
Knowledge	8	Compute Multiple Integrand	2.2.	M	2
Knowledge	11a	Jacobian	N/A.	F	1

Note. M = Multiple Choice; F = Final Answer; W = Written

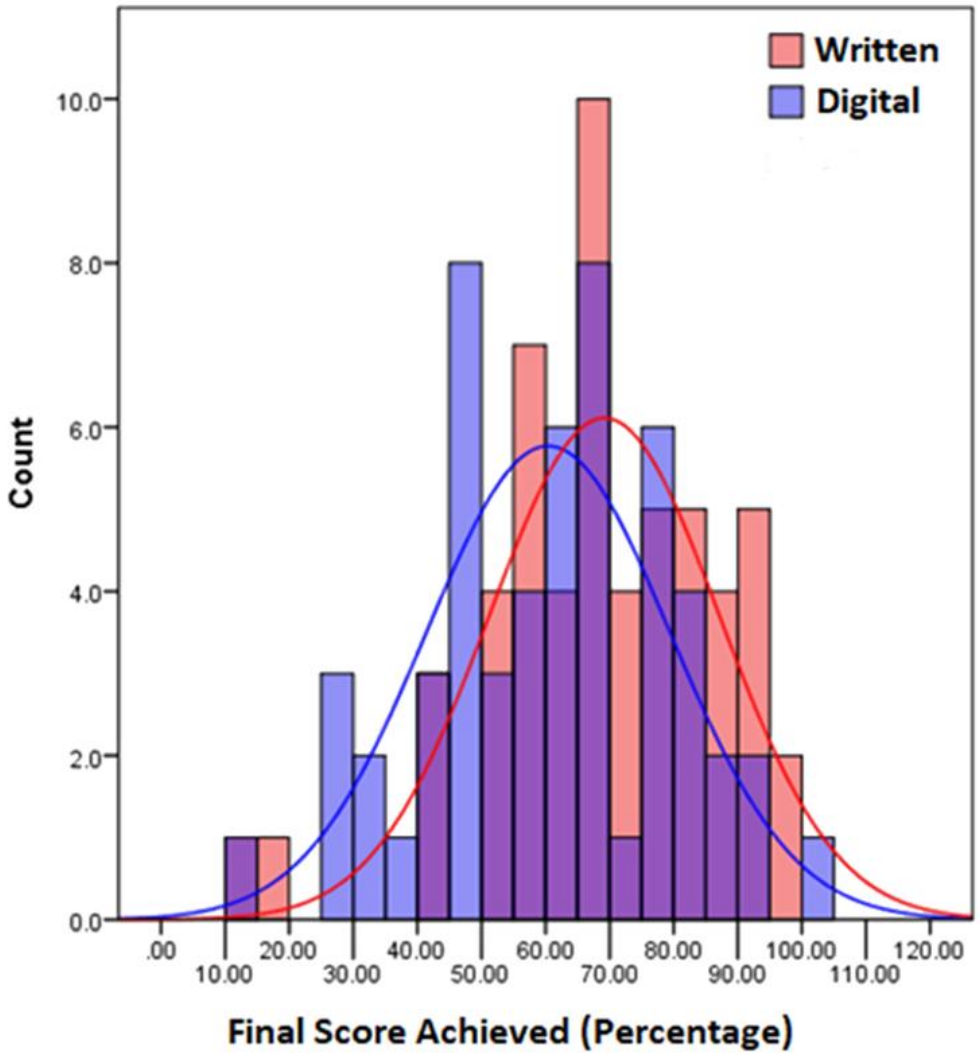
Results sub-question 2:
To what extent can digital testing questions create the expected distribution of students according to their mathematical ability?

Results for sub-question 2 from the 2016 pilot.

In the pilot from 2016, it was checked if there is a significant difference in the mean test marks between the paper-based results against the equivalent digital questions using a paired sample t-test. The average mean in the written version was 68.4% (*SD* = 17.9), whilst that of the digital version of the same exam was 59.7% (*SD* = 19.0). The paired sample T-test revealed a significant difference in the means, with a difference of 8.68% in the means, with $p < .001$ (two-tailed). Graphically, this can be seen represented in Diagram 1, where the distribution curve of digital testing questions in the 2016 to the left of the equivalent questions answered in a written format. The paper-based questions are depicted in red, whilst the equivalent digital questions are shown in blue.

Diagram 1

Comparison of the Final Percentage Achieved when Assessed by Hand (written) and Assessed Digitally in the 2016 Pilot.



Analysis of both written and digital equivalent questions are represented in table 7. The table displays the results of P-values, and the Cronbach Alpha if the item is deleted. In order to detect items that are possibly of poor quality, certain criteria were set, and values bolded in table 7. If the difference in p-value between the written and the digital questions were more than .1, these values were bolded. The Cronbach's Alpha for the written test was found to be .68 and for the digital test .58. Values higher than these in the Cronbach if deleted column, were bolded. As a result, from the 2016 pilot, questions 1a, 1 c, 2, 6b, 7 and 8a all have values that are bolded and should be further investigated in conjunction with other sub-questions in this research.

Table 7

P-value and Reliability of Items for both the Written and Digital Methods of Assessment for the 2016 Pilot.

Question	Max Score	Score breakdown	P- Values		Cronbach if deleted	
Final Answer			W	D	W	D
Q1a_Velocity	1	3 of 0.33	.96	.96	**.68**	.57
Q1b_Acceleration	1	3 of 0.33	.86	.87	.68	.57
Q1c_Speed	1	1	.93	.88	**.69**	.58
Q1d_UnitVector	1	3 of 0.33	.67	.63	.68	.57
Q2_DirDerivative	3	3	**.54**	**.27**	.66	**.61**
Q3_LocalMaxMinSaddle	6	3 of 2	.82	.76	.60	.48
Q4_Lagrange	5	2 of 2.5	.61	.58	.65	.50
Q6a_VertCrossSection	2	4 of 0.5	.83	.74	.65	.55
Q6b_HorizCrossSection	2	4 of 0.5	**.76**	**.66**	.66	.55
Q7_VolumeCylinder	6	6	**.45**	**.29**	**.71**	**.64**
Q8a_Jacobian	3	3	**.54**	**.39**	.63	.51
Q9_Non-LinearTransform	2	4 of 0.25 and 1 of 1	.55	.61	.66	.53
Multiple Choice						
Q5_ IntegrationSketch	1	1	.80	.82	.68	.57
Q8b_ImageTransformation	2	2	.87	.93	.67	.56
Max	36	36	.68	.60	.68	.58

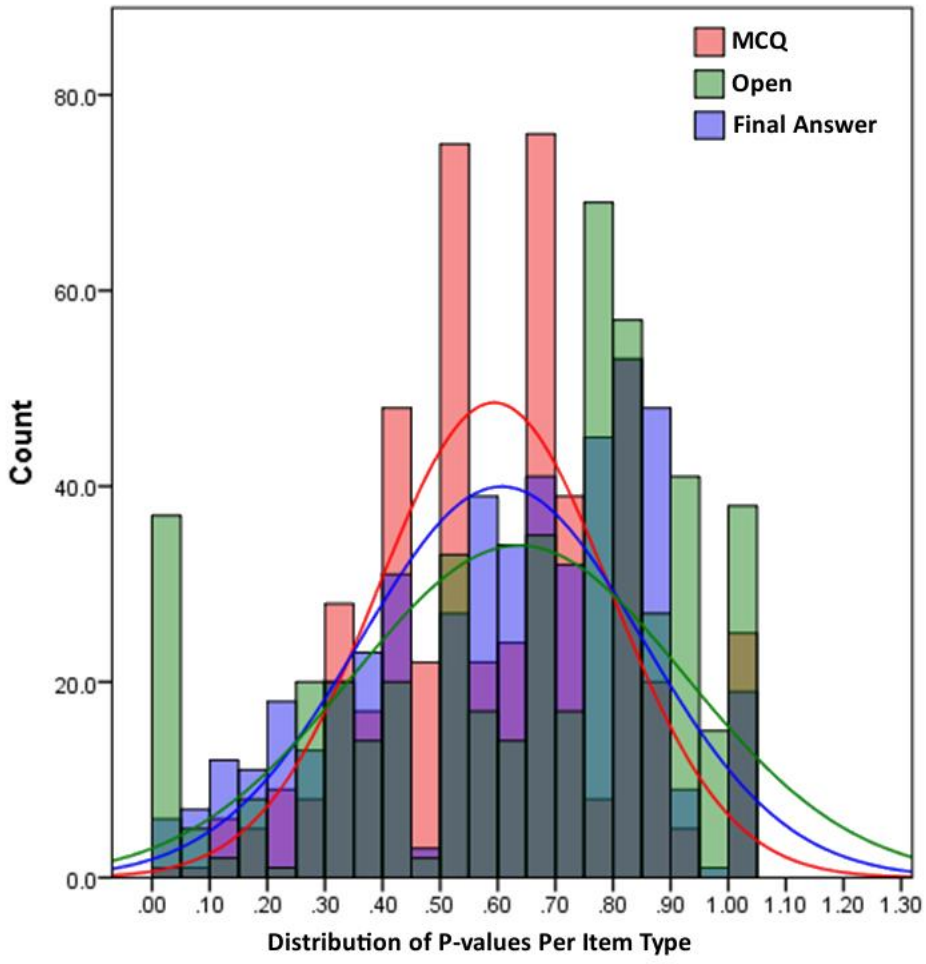
Results for sub-question 2 from the 2017 pilot.

In the 2017 pilot, the exam consists of multiple-choice questions, final answer questions and open written questions. The P-value of the open questions have a mean of .64 and *SD* = .29; The p-value of the multiple choice questions (MCQ) have a mean of .59 and *SD* = .20 , and the p-value for the final answer questions has a mean of .61 and *SD* = .25.

Three paired samples t-test was done between the three variables to discover where there might be a statistical difference in the means of the different item types. Significance in means was done with a two-tailed test. Significant results include Open questions and MCQ had showed a different of .04 in the mean with a significance of p =.001; Open questions and Final answer showed a difference in the mean of .03 with a significance of p = .005. The different p-values of each item types and their distribution is displayed in diagram 3, where the slight, but significant differences in the mean can be seen.

Diagram 2

The P-values of the Three Item Types of the 2017 Exam Compared.



Since there seems to be a similar distribution of the three question types in the 2017 pilot, however, the open question have a large first and final bar, due to

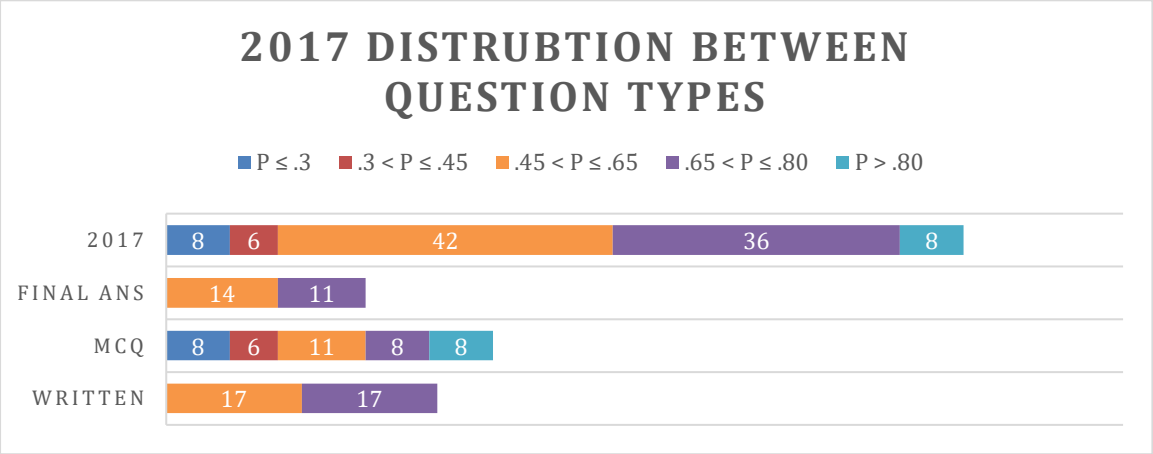
question 10. The correlation between the three item types was investigated using a pearson correlation, in order to see if one part of the exam can predict the other. The correlations are significant at a two-tailed significance, but all are below a .7 correlation.

Table 8
Pearson Correlation Different Items Types in 2017 Exam

	Open	MCQ	Final Answer
Open			
MCQ	.39**		
Final Answer	.63**	.39**	

Note. Open refers to questions answered on paper, MCQ is for Multiple Choice Questions and "Final Answer" refers to questions that are only judged based on their final answer.

Diagram 3
Distribution of p-values Across Different Item Types



Analysis of all three item types are displayed in table 9. Table 9 displays the results of P-values, and the Cronbach Alpha if the item is deleted of each item. In order to detect are possibly of poor quality, certain criteria was set and values bolded in table. If the p-value was less than .3 or more than .85, it was bolded. The Cronbach's Alpha for the written test was found to be .69 Values higher than these in the Cronbach if deleted column, were bolded. As a result, from the 2017 pilot, questions 2 and 8 all have values that are bolded and should be further investigated in conjunction with other sub-questions in this research.

Table 9
P-values, Correlation and Reliability for 2017 Exam.

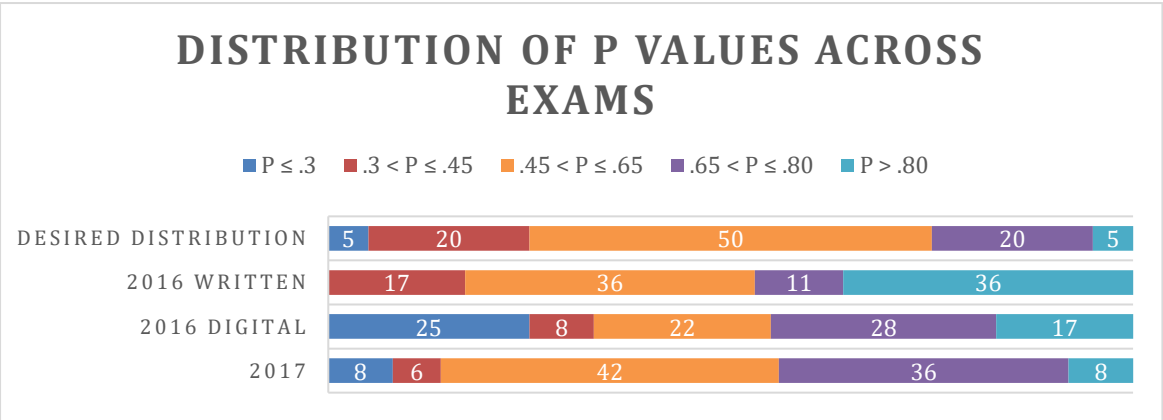
Question	Possible Score	Score breakdown	Mean	SD	P-value	Cronbach if deleted
Multiple Choice						
Q2_ChainRule	3	3	0.79	1.3	**.27**	**.71**
Q3_GradientVector	2	2	0.78	1.0	.40	.68
Q4_TangentLine	2	2	1.19	1.0	.60	.69
Q5_CriticalPoints	3	4 of 0.75	2.33	1.0	.79	.67
Q7_IntRegion	1	1	0.82	0.4	.83	.69
Q8_MultipleIntegrand	2	2	1.74	0.7	**.88**	.68
Q11b_IntUnderTrans	2	2	1.25	1.0	.64	.68
Final Answer						
Q1_Vectors	2	4 of 0.5	1.24	0.7	.63	.68
Q9_NewIntLimits	3	1 of 0.5 and 6 of 0.25	2.03	1.2	.70	.66
Q11a_Jacobian	1	1	0.73	0.4	.78	.68
Q12_NewIntegrand	3	4 of 0.5 and 1 of 1	1.47	1.0	.52	.65
Open answer						
Q6_OptimumLagrange	6	6	3.72	1.7	.64	.64
Q10_TripleIntCylinder	6	6	3.92	2.4	.70	.63

Note. As mentioned in text, Cronbach's Alpha was .69

Results for sub-question 2 by comparing 2016 and 2017 pilot.

For an overview of the two pilots and the spread of p-values in each exam, diagram 4 has spread the questions types across different p-value categories.

Diagram 4
Distribution of p-values across different exams.



Results sub-question 3:
How is overall and item-wise discrimination effected by digital testing?

Results for sub-question 3 from the 2016 pilot.

Table 10 shows the Corrected Item-Total Correlation (CITC), the extreme two-group discrimination displaying the difference in p-value between the top 25% and bottom 25%. In CITC, values less than or equal to .3 were bolded. In the two-group discrimination, values below .3 were bolded. In sub-question 2, 1a, 1 c, 2, 6b, 7 and 8a, were raised for concerns. These same items appear again, and in addition also 1b,1d, 5 and 8b.

Table 10
Discrimination of Both Written and Digitally Assessed Items in 2016 Exam

Question	Max	Score breakdown	Corrected item-total correlation		Two group Discrimination	
Final Answer			*W*	*D*	*W*	*D*
Q1a_Velocity	1	3 of 0.33	**.09**	**.30**	**.07**	**.10**
Q1b_Acceleration	1	3 of 0.33	.33	.35	**.20**	**.18**
Q1c_Speed	1	1	**-.06**	**.10**	**-.03**	**.08**
Q1d_UnitVector	1	3 of 0.33	**.23**	**.21**	.42	.52
Q2_DirDerivative	3	**3**	.31	**-.02**	.51	**.29**
Q3_LocalMaxMinSaddle	6	3 of 2	.63	.48	.46	.45
Q4_Lagrange	5	2 of 2.5	.42	.43	.48	.56
Q6a_VertCrossSection	2	4 of 0.5	.48	.33	.35	.33
Q6b_HorizCrossSection	2	4 of 0.5	.36	.31	**.28**	**.30**
Q7_VolumeCylinder	6	6	**.28**	**.14**	.54	**.24**
Q8a_Jacobian	3	**3**	.50	.43	.72	.68
Q9_Non-LinearTransform	2	4 of 0.25 and 1 of 1	.41	.59	.39	.39
Multiple Choice						
Q5_ IntegrationSketch	1	1	**.21**	**.25**	**.22**	**.22**
Q8b_ImageTransformation	2	2	.31	**.27**	**.30**	**.14**
Max	36	36			.45	.38

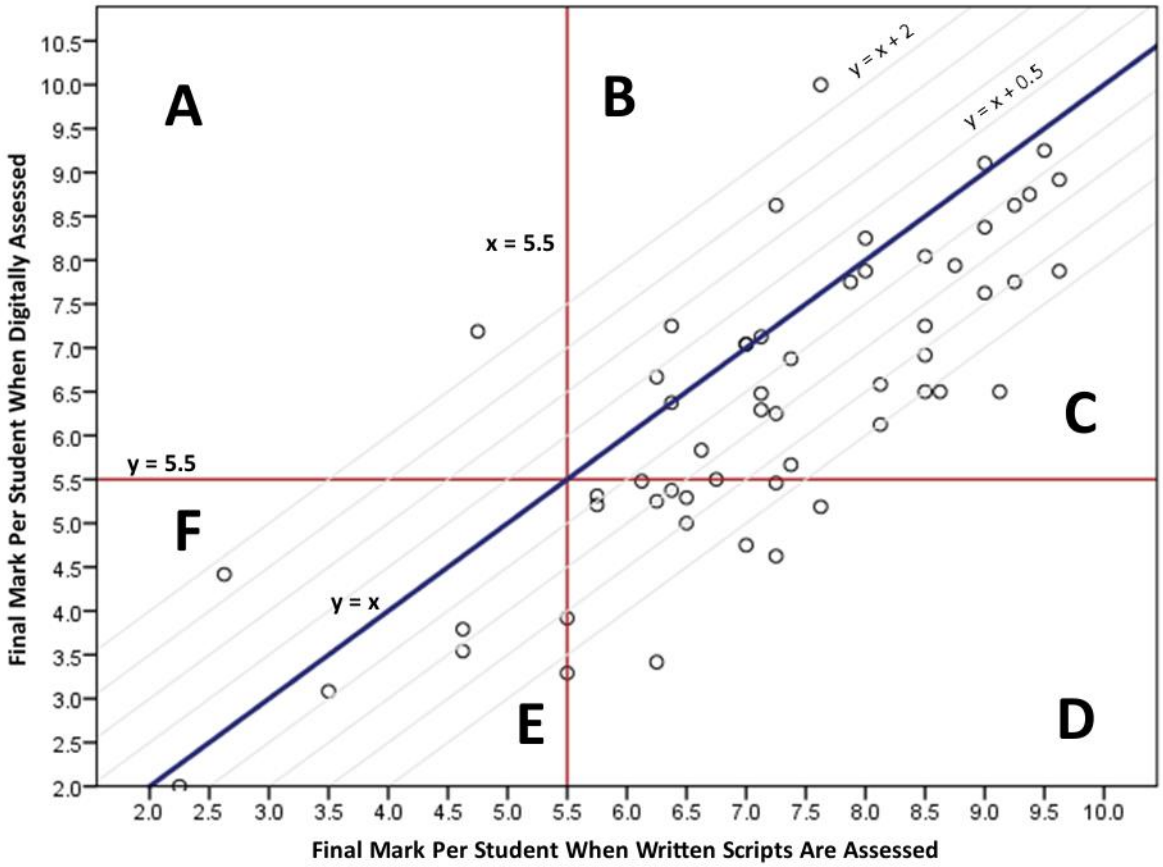
Diagram 5 shows a scatterplot of the 2016 pilot exam. On the x axis are the marks achieved in the written exam and in the y axis the marks achieved for the same student in the equivalent digital exam. The diagram has various reference lines as described in the Data Analysis section and in the notes below the graph. Between the reference lines are sections labelled A to F. These sections make use of the y =x, x = 5.5 and y = 5.5 lines to show the following: Section A contains students that would fail in a written exam but pass when the exam was digitally graded (n =1). Section B contains students that are passing in the written exam and gaining even higher marks when digitally graded (n = 7). It is noted

that there are two students (n = 2) precisely on the y = x line. They are not counted in either the B or C section. The C section contains students that are passing with the written exam, but they are doing less well when assessed digitally (n = 26). Section D contains students that would pass in a written exam but would fail when assessed digitally (n = 14). Section E and F are counted together, with students that are failing in a written exam, and would also fail in an equivalent digital exam (n = 5).

The grey reference lines show by how much students are affected by the change in exams. Irrelevant of section A to F, a distribution of scores can be seen in table 11.

Diagram 5

Scatter Plot of Written Exam Plotted Against the Digital Exam Equivalent



Note. The blue line, y = x , is where a student would get the same mark for an exam, regardless if they were assessed digitally or by hand. Grey lines differ by 0.5 from the blue y=x line.

Table 11

The Gain or Loss of Marks of Students When Written Exams are Assessed Digitally

	N losing marks	Cumulative total	N gaining Marks	Culminative total
x >2	7	7	2	2
1.5 < x ≤ 2	8	15	1	3
1 < x ≤ 1.5	7	22	1	4
0.5 < x ≤ 1	14	26	1	5
0 < x ≤ 0.5	8	44	4	9

In the 2016 pilot, the examinees were split into three groups, low represented the bottom 25% based on the written exam, and the high group the top 25% based on the written exam. The difference between P-values, Written minus Digital, were done for each group and is represented in table 12. A paired sample T-test was conducted on the difference between the written and digital exams to see if there is a significance change in the means of the p-values per academic group per question. In the table significant values are bolded, indicating that certain questions might be responsible to the change in marks between the digital and written equivalent exam. Table 12 indicated that Questions 2 , 6a, 6b, 7, 8a need to be investigated further for the reason behind the decree in marks from the written component to the digital component, whilst question 9 might be the reason for a gain in marks from written to digital.

Table 12

Difference of P-value by Group in 2016 Exam Per Item, Written – Digital.

Question	Diff overall P	Diff Low P[a]	Diff Mid P[b]	Diff High P[c]
Multiple Choice				
Q1a_Velocity	.00	0.02	0.00	0.00
Q1b_Acceleration	–.01	-0.03	0.03	-0.04
Q1c_Speed	.05	0.08	0.08	0.00
Q1d_UnitVector	.04	0.00	0.13	-0.02
Q2_DirDerivative	**.27****	**0.16****	**0.28****	**0.37****
Q3_LocalMaxMinSaddle	.06	0.07	0.06	0.06
Q4_Lagrange	.03	0.05	0.02	0.03
Q6a_VertCrossSection	**.90***	0.04	0.11	0.13
Q6b_HorizCrossSection	**.10****	0.12	0.07	0.13
Q7_VolumeCylinder	**.16****	0.11	0.06	**0.41****
Q8a_Jacobian	**.15****	0.11	**0.21****	0.13
Q9_NonLinearTransform	–**.06***	-0.02	**-0.13****	-0.04
Final Answer				
Q5_ IntegrationSketch	–.02	.03	-0.08	0.00
Q8b_ImageTransformation	–.06	–.17	0.00	0.00

Note. [a]n=19, of students with written mark up to 6.5; [b]n=19; [c]n=18 of students with written mark of more than 8.1. Paired Sample T-test with 95% Confidence Interval Percentage. *significance at 0.05 level; **significance at 0.01 level

Results for sub-question 3 from the 2017 pilot.

In table 13 the Corrected Item-Total Correlation (CITC) and the two-group discrimination displaying the difference in p-value between the top 25% and bottom 25% can be seen. In CITC, values less than or equal to .3 were bolded. In the two-group discrimination, values below .3 were bolded. In sub-question 2, questions 2 and 8 were marked for investigation. Here they are too, in addition to questions 3, 4, 7 and 11b.

Table 13
Discrimination of Items in 2017 Pilot

Question	Possible Score	Mean	Two Group Disc	Corrected Item-Total Correlation
Multiple Choice				
Q2_ChainRule	3	0.79	.33	**.12**
Q3_GradientVector	2	0.78	.49	**.23**
Q4_TangentLine	2	1.19	.39	**.18**
Q5_CriticalPoints	3	2.33	.36	.34
Q7_IntRegion	1	0.82	.31	**.27**
Q8_MultipleIntegrand	2	1.74	**.28**	**.29**
Q11b_IntUnderTrans	2	1.25	.50	**.29**
Final Answer				
Q1_Vectors	2	1.24	**.29**	**.28**
Q9_NewIntegrationLimits	3	2.03	.60	.43
Q11a_Jacobian	1	0.73	.38	.41
Q12_NewIntegrand	3	1.47	.52	.50
Open answer				
Q6_OptimumLagrange	6	3.72	.44	.49
Q10_TripleIntCylinder	6	3.92	.80	.58

In table 14, distractors from the 2017 were examined for the percentage they were chosen. Distractors chosen by more than 5% were regarded as good and bolded. Due to the many figures in this table, we have rounded to 0 decimal places, except where the number is below 1%. Also in table 14, RAR values for distractors in the 2017 exam were analysed. Question 5 is not a typical multiple-choice question, as it had four critical points that were given, and then for all four points, a choice could be made between four names for each point. Values lower than -2 were bolded showing effective distractors, and positive RAR values in the distractors were bolded to highlight troublesome distractors. In question 2, a seems effective, whilst g is troublesome.

Table 14

Analysis of Distractors of MCQ items in the 2017 Pilot

	Percentages per distractor in 2017 Pilot									RAR values per distractor in 2017 Pilot							
Q	a	b	c	d	e	f	g	h	Nothing	a	b	c	d	e	f	g	h
Q2	2	3	2	26	2	4	**54**	**6**	0.8	**–.24****	–.06	–.03	.30**	–.16**	–.10*	**.02**	–.17
Q3	5	3	**8**	**19**	39	**6**	**16**	5	0.8	–.16**	–.05	–.11*	–.08	.36**	**–.22****	–.02	–.05
Q4	**17**	**7**	4	1	**7**	2	60	2	1.2	–.07	–.10*	–.15**	–.03	–.15**	–.13**	.31**	–.05
Q5_1	75	**14**	**9**	1					0.8	.34**	**–.22****	–.16**	–.10*				
Q5_2	4	**8**	**8**	80					0.8	–.18*	**–.21****	–.17**	.37**				
Q5_3	4	**10**	78	**8**					0.8	–.12**	–.17**	.32**	–.18**				
Q5_4	**15**	77	**6**	2					0.8	**–.27****	.36**	–.12**	–.12*				
Q7	**7**	0.2	2	3	2	2	82	1	0.6	–.09	–.05	–.13**	**–.27****	–.11*	–.05	.31**	–.09
Q8	1	1	87	3	0.4	0.6	**4**	2	0.6	–.11*	–.06	.37**	–.17**	–.16**	–.07	–.14**	**–.21****
Q9	24	75							1.0	**–.43****	.46**						
Q11b	3	5	3	63	**18**	**8**			1.2	–.16**	–.14*	–.18**	.41**	–.18**	–.10*		

Note. Correct Answers are underlined and not in boldface. >5% chosen indicated good distractors, therefore is in boldface. RAR values below –.2 indicate good distractors for discrimination and are in boldface. RAR values above 0 are considered bad and are in boldface. **Correlation is significant at the 0.01 level (2-tailed); *Correlation is significant at the 0.05 level (2-tailed)

Results sub-question 4:
What is the current state of acceptance of digital testing calculus amongst first year engineering students?

Results for sub-question 4 from the 2016 pilot.

Table 15 contains the number of respondents (N), the mean and the Standard Deviation (SD) for the survey questions taken at the end of the 2016 exam. Questions were based on a five point Likert scale from 1 = totally disagree to 5 = totally agree, except for question *1 where 1 = 0-5 min, 2= 5-10min, 3=10-15min, 4=15-20min, 5 >=20min*.

Table 15
Evaluation Reponses in the 2016 Exam

Measure	N	Mean	SD
Q2_ExpMLP	55	3.71	0.79
Q3_InputMLP	54	3.98	1.07
Q4_DigiOnlyFair	55	1.58	0.79
Q5_DigiOnlyAdv	55	1.95	1.02
Q6_DigiOnlyNeat	55	2.40	1.12
Q7_CheckAns	55	3.05	1.10
Q8_GoodWay	55	2.40	1.07
Q9_FairWay	55	2.55	1.02
Q10_OwnDevice	55	4.20	0.99
Q11_FraudProof	55	3.09	1.19

Note. Q1 was excluded due to being irrelevant and in minutes
Q2 to Q11 consisted of scale 1 = totally disagree to 5 = totally agree.

Evaluation questions from the 2016 exam were evaluated per point to discover trends in the answers given by students. This table can be found in the appendix 12.

The final variables and the means of the variables that make up the 2016 digital acceptance construct is presented in table 16.

Table 16
Evaluation Reponses of all Students in the 2016 Exam.

Measure	N	Mean	SD
Q4_DigiOnlyFair	55	1.58	0.79
Q8_GoodWay	55	2.40	1.07
Q9_FairWay	55	2.55	1.02
DigiAcceptance2016	55	2.18	0.76

Results for sub-question 4 from the 2017 pilot.

Table 17 and table 18 contains the number of respondents (N), the mean (Mean) and the standard deviation (SD) for the survey questions taken at the end of the 2017 exam. Table 17 contains seven questions measuring the opinions students have regarding digital testing, from both the Electrical Engineering Students and the Paper-based students combined.

Table 17

Evaluation Reponses of all students in the 2017 exam.

Measure	N	Mean	SD
Q1_GoodWay	365	2.96	1.17
Q2_FairWay	373	2.79	1.10
Q3_MCQEasier	362	2.84	0.98
Q4_Ratio Right	366	2.93	1.10
Q5_RatioPrefer	330	3.37	1.01
Q6_TradBetterScore	355	3.06	0.94
Q7_TutGoodPrep	351	1.25	0.43

Note. Q1-Q4 and Q6 were on the scale: 1 = totally disagree to 5 = totally agree. Q5 was scale (MCQ – W) 1 = 100 - 0, 2 = 75 - 25, 3 = 50 - 50, 4 = 25 - 75, 5 = 0 – 100. Q7 was 1= yes, 2 = no.

Table 18

Evaluation Reponses in the 2017 Exam, of Electrical Engineering Students.

Measure	N	Mean	SD
Q8_InputMLP	44	3.84	1.18
Q9_WIFIGood	44	4.73	0.50
Q10_NoTechIssues	44	4.55	0.88
Q12_CheckAns	43	2.67	1.44
Q13_ShowScore	44	2.52	1.25
Q14_FraudProof	44	2.98	1.02
Q15_PreferDigitalExam	44	2.32	0.88

Note. Questions were on the scale: 1 = totally disagree to 5 = totally agree.

Table 19 contains detailed descriptions for questions 4 and 5. It is noted that when the EE students chose that they are dissatisfied with the ratio in the exam, they all chose a number indicating that they wanted less digital testing in the exam. Through the detailed table, 259 of the 330 (78.5%) students that filled in question 5, wanted 50% or less of the exam to be digital. When compared to the entire exam group then it could be said that, 259 of 492 students (52.6%) would like the paper to consist of 50% testing questions or less. On the other hand, only 13.6% of the 330 students would like no digital testing whatsoever, which is 9% of the 492 students.

Table 19
Detailed evaluation Reponses of all Students in the 2017 Exam.

Measure	N	1	2	3	4	5	Mean	SD
Q4_RatioRight_Total	366	25	**130**	82	102	27	2.93	1.10
Q4_RatioRight_EE	44		**31**		13		2.60	0.92
Q4_RatioRight_Other	322	25	**99**	82	89	27	2.98	1.11
Q5_RatioPrefer_Total	330	5	66	**110**	100	49	3.37	1.01
Q5_RatioPrefer_EE	30	0	0	**14**	12	4	3.66	0.71
Q5_RatioPrefer_Other	300	5	66	**96**	88	45	3.34	1.03

Note. Question 4 in the EE exam was limited to yes and no – this was converted to agree and disagree. Only those that disagreed, could answer question 5, whilst this was not the case for the "other" group. Q4: 1 = totally disagree to 5 = totally agree.
Q5 (MCQ – W): 1 = 100 - 0, 2 = 75 - 25, 3 = 50 - 50, 4 = 25 - 75, 5 = 0 – 100.

Table 20 presents the final digital acceptance variable for 2017, with a mean (Mean) of 2.83 and standard deviation (SD) of .836.

Table 20
Digital Acceptance Among all Students in the 2017 Exam.

Measure	N	Mean	SD
Q1_GoodWay	365	2.96	1.17
Q2_FairWay	373	2.79	1.10
Q5_ReCRatioPrefer	330	2.63	1.01
Q6_ReCTradBetterScore	355	2.94	0.94
DigiAcceptance2017	280	2.83	.84

Note. Q1 to Q4; Q6, was scale: 1 = totally disagree to 5 = totally agree.
Q5 was scale (MCQ – W) 1 = 100 - 0, 2 = 75 - 25, 3 = 50 - 50, 4 = 25 - 75, 5 = 0 – 100. Q7 was 1= yes, 2 = no.

Results for sub-question 4 from comparing the 2017 and 2017 pilots.

Comparing two digital acceptance variables from 2016 and 2017, a .65 increase on average from 2016 to 2017 can be seen, where the variable DigiAcceptance2016 has mean = 2.18 and *SD* = .756 and the variable DigiAcceptance2017 has mean = 2.83 and *SD* = .836. An independent sample t-test was conducted with the results from 2016 and 2017. (2016 n =55, 2017 n = 277). Equal variance was passed with p = .226 and the difference in means is significant with p < .001. (two-tailed significance).

A one way ANOVA is conducted to see if there is a difference in the means between high, low and medium academic achievers in their acceptance in digital acceptance. In both 2016 and 2017 the low, medium and high groups were split based on their mark out of 36. Low had a mark up to and including 21 (causing the student to get a "6"), whilst the high group is a mark of 27 or higher (gaining the student exactly a mark of "7.5" or higher), and the middle group is in between. In 2016 this split done with the written exams. There was a minimum of 13 per

group. The 2016 written group tested positive for an equal variance with p =.156 for the groups, and tested negative for a difference in the means of digital acceptance between the three academic groups with p = .153.

The 2017 results tested positive for equal variance with p = .922 but the difference in means of digital acceptance was rejected with a p = .610.

In the 2016 and 2017 pilot, student could fill out written answers to the question "Suggestions". These open-answers in both pilots were searched for key words, from which a coding scheme was developed. This coding scheme can be seen in appendix 11 . The responses were searched and tallied in Table 21.

Table 21
Categories for the Open Answer Questions in 2016 and 2017

Code	*Label*	*2016*	*2017*
1	Digital testing is worse/disliked and/or is stressful	7	5
2	Digital is unfair; Cannot show what you know	15	6
3	One error result in losing all marks	10	7
4	Method is more important than answer	9	2
5	More steps digitally	3	2
6	Comments on design/ratio of the exam	2	4
7	Positive or OK remark	1	3
8	Practical suggestions for exam venue/layout	3	5
9	Human checking also	3	0
10	Chromebook for Maths is laborious	7	4
11	Time consuming	2	1
12	Fraud concerns	5	0
13	Good for diagnostic exams	2	0
98	Other	3	4
99	Don't know/Other	0	1

The top three scoring feedback that was given, were concerns in both 2016 and 2017 of Digital testing being disliked and stressful; that digital testing is unfair as you cannot show/get points for what you know and by making one error, you lose all the marks for one point. These top three remarks closely tie into the 4th highest comment in 2016, which is that the method/calculation is more important than the final answer. Comments in 2017 contained questions about the ratio of the exam which included having more written questions, wanting a button that marks a question as filled in when left blank and a comment that examiners should avoiding the cascade of wrong answers. The 2017 open answers also contained more positive remarks which included remarks such as "A step into the future"; "worked fine for me" and a contradictory "I like this way of testing, but it is unfair". Noteworthy suggestions also include a remark which says, "I would like to quote our lecturer: What Eve did to paradise, is what multiple-choice does to maths education." As well as a suggestion from another student that tests should also

be done on paper and students can then request a remarking of the paper-based exam.

Results for sub-question 4 from the focus group.

In order to get the view from the lecturers at the University of Twente, a focus group was conducted with 6 lecturers. A full summary of the focus group can be seen in appendix 7. Important points that came out of the discussion are:

Advantages

- Saves a lot of time, but only with big groups
- Statistics gathered are really useful with item selection and item order for future exams
- Digital testing also gives access to diagnostic testing, which should be done more during courses
- Digital testing is not lenient towards sloppy work, which could be advantageous if/when used correctly.
- Digital testing is good for basic questions, and it is expected that 80% of questions for summative exams can be handled digitally
- MCQ gives more information regarding misconceptions to the teacher then final answer with current systems

Disadvantages

- more time is spent on writing the items
- MCQ is limited in what you can ask due to reverse engineering, which meet different educational goals then desired
- students gets punished for small mistakes – which is regarded unfair and is hard to justify to both staff and students
- Final answer has various concerns – one of which is the all or nothing approach which appears to be very shallow assessment
- Basic questions are open to all or nothing approach in final answer, but setting good questions like that is really hard
- Expectation management and careful item management might help, but cannot prevent all problems with digital testing
- Digital testing might save time, but if it does test what you want to test, the feeling is "so what if it saves time"

Neutral comments

- It should be made clear the philosophy behind digital testing is not so that small mistakes should be punished.
- When introducing a new digital item, it is unknow how students will react to it – a pilot is needed
- For calculus an item bank is available
- Checking process and answer should be a mandatory learning goal.
- More than a sample exam should be done to prepare students for digital exams ,such as learning to check answers to not loose points.

Other comments regarding what to keep in mind when pursuing digital testing and future possibilities is covered in appendix 7.

As an email was sent out inviting staff to comment if they are not able to attend, one lecturer with more than 35 years' experience in teaching made use of this opportunity, with the following email:
Direct quotations include:

- I understand that increased student numbers seem to generate a need for digital testing, but in my opinion the students' knowledge that we can test this way is rather limited.
- At best one could add a few multiple-choice questions trying to check whether they understood certain concepts
- So-called "do exercises" where the student is to compute a solution and input the result in a digital system are unsatisfactory because faults can arise in many ways.
- I usually include some multiple-choice tests in exams for [….] students and this works well because I ask them to motivate their answer in one or two sentences. This does check knowledge quite well and is still easy to mark.
- Neither did it save a lot of time since exam preparation was rather expensive:
- Quite a few students seem to like mylabs+, so digital testing is good for training, but not for an exam, I would say.
- For an average exam I usually spend 10 min per student (at least after checking the first 10 or so).

5.3 Analysis of Results

Due to many items needing further analysis, all four sub-questions will be combined and analysed using the earmarked items. Two tables will assist to summarise the information. First the 2016 results will be discussed with the use of table 22 and then discussed how these were applied in the 2017 pilot, with the use of table 23. Sub-question 4, discussing student and staff acceptance will be discussed more at the end of these two sections.

Table 22

Analysis of Results from the Digital Component of 2016 Exam

Question	Score	P- Values	Mode	Cronbach if deleted	CITC	Two Group Discrimination	Blooms	Course Goals
Troublesome items								
Q2_ DirDerivative	3	.27 [a]	F	.61	**-.02**	**.29 [a]**	Knowledge	1.3.; 1.4.
Q6a_ VertCrossSection[a]	4 of 0.5	.74	F	.55	.33	.33	Application	2.1.
Q6b_ HorizCrossSection	4 of 0.5	.66 [a]	F	.55	.31	.30	Application	2.1.
Q7_ VolumeCylinder	6	.29 [a]	F	.64	.14	.24 [a]	Analysis	2.3.; 2.4.
Q8a_ Jacobian	3	.39 [a]	F	.51	.43	.68	Knowledge	N.A.
Troublesome due to low/change in discrimination								
Q1a_Velocity	3 of 0.33	.96	F	.57	.30 [a]	.10	Knowledge	1.1.
Q1b_Acceleration	3 of 0.33	.87	F	.57	.35	.18	Knowledge	1.1.
Q1c_Speed	1	.88	F	.58	.10	.08 [a]	Knowledge	1.1.
Q1d_UnitVector	3 of 0.33	.63	F	.57	.21	.52	Knowledge	1.1.
Q5_ IntegrationSketch	1	.82	M	.57	.25	.22	Comprehension	2.1.
Q8b_ImageTransformation	2	.93	M	.56	.27	.14 [a]	Knowledge	2.4.; 2.1.
Good items								
Q3_ LocalMaxMinSaddle	3 of 2	.76	F	.48	.48	.45	Comprehension	1.4.; 1.10.
Q4_ Lagrange	2 of 2.5	.58	F	.50	.43	.56	Comprehension	1.12.; 1.4.
Q9_ NonLinearTransform	4 of 0.25 and 1 of 1	.61	F	.53	.59	.39	Synthesis	2.4.; 2.1.

Note. Cronbach Alpha for digital component was .58. As these the digitally assessed marks, those with [a] indicate where the difference is more than .10. [a]6a is a special case as it only has a .09 difference in p-value with digital, but the difference is significant w.r.t. the written version.

Analysis of Results from the 2016 pilot.

The research questions each looked at different components to discover to what digital testing can occur in summative Mathematics exams. Sub-question 1 looked at Bloom's Taxonomy and the course goals of items, sub-question 2 at the p-values of items and sub-question 3 at the discrimination of items, in terms of the difference in groups and how each items discriminates and sub-question 4 at the digital acceptance of staff and students. In the first sub-question, the analysis did not reveal anything significant about course goals, except that they were adequately covered. Contrary to expectations both Analysis and Synthesis from Bloom's Taxonomy appeared in the results. During the coding of these items with the content expert, the expert made a special mention about question 2, being "This question has 4 procedures before getting to the answer, students will struggle to get to a final answer without a mistake." In sub-question 2 the analysis of p-values highlighted a few troublesome questions, namely question 2, 6b, 7 and 8. All of these has .10 of less in the p-value than the written component. This is troublesome in terms of criterion-referenced testing, which is concerned with students getting the same marks across tests. Sub-question 3 showed a scatterplot where 22 of the 55 (40%) students were having a difference of 1 mark or more when their results were assessed digitally. This could be narrowed down to few questions causing this, namely again questions 2, 6a, 6b, 7 and 8a. Question 2 in particular had a negative CITC, showing that it was unreliable and testing a construct different from the rest of the exam. It is perhaps then no surprise that when digital acceptance of these students was assessed, it was 2.18 on a 5 point Likert scale, indicating that they are not in favour of having their exams assessed digitally. Most of the written comments were concerns regarding making one typing error, or one calculation error and then losing all their marks. This is similar to the validity concern mentioned in the theoretical framework – if one mistake (a "." instead of a "," as in question 8a) causes the losing of marks, it should be investigated if the item really then testing the construct of Mathematical ability.

In table 22, all the characteristics of the items across the sub-questions are presented. It is not sufficient to say that it is a particular type of Bloom's taxonomy that causes questions to be less suitable for digital testing. In fact, Synthesis, which is the highest of them all, is labelled to be a good item. There is also no pattern in terms of the course goals. Thus, the mark breakdown is considered to be the greatest predictor of digital testing success. Questions 2, 7 and 8a all have 3 or more marks, but then 6a and 6b have the 0.5-mark increments, the same as the "good" question 9. Some insight reveals that with question 6a and 6b, the difference between the written and digital versions come in, because in the digital versions they first have to choose between $dydx$ and $dxdy$ and then fill in the integrand limits. Choosing the wrong order, results in 0 marks, no matter what the student fills in for limits. This is not the case in the written answers. Thus, there is a weakness in the validity of this question, as in the written question, a student can still show with the correct limits an understanding of what the question

is testing. Question 8a was discussed with a content expert and it was unclear why this question was performing badly digitally, as it is a straightforward question. However, three marks seemed high and this, or the "," versus "." Issue could be the reason. Question 2 and 7 consisted of too many steps that were then assessed by one final answer of more than 3 marks.

As a conclusion, changes that were implemented in the 2017 exam was that questions should have simplistic calculations that lead to the final answer (unlike question 2) and also be of less marks, avoiding 3-mark final answer questions, and having more questions such as question 9 which has 0.5-mark increments.

Table 23

Overview of results from all sub-questions in 2017 pilot

Question	Score	P-value	Mode	Cronbach if deleted	Two Group Discrimination	CITC	Blooms	Goals
Troublesome items								
Q2_ChainRule	3	.27[a]	M[a]	.71[a]	.33	.12[a]	Application	1.2
Q4_TangentLine	2	.60	M[a]	.69	.39	.18[a]	Application	1.5
Borderline items due to low CITC								
Q1_Vectors	4 of 0.5	.63	F	.68	.29[a]	.28[a]	Knowledge	1.1; 1.2
Q3_GradientVector	2	.40	M[a]	.68	.49	.23[a]	Knowledge	1.4; 1.5
Q8_MultipleIntegrand	2	.88	M	.68	.28[a]	.29[a]	Knowledge	2.2
Q7_IntRegion	1	.83	M	.69	.31	.27[a]	Application	2.1
Q11b_IntUnderTrans	2	.64	M	.68	.50	.29[a]	Application	2.1; 2.4
Good items								
Q5_CriticalPoints	4 of 0.75	.79	M	.67	.36	.34	Knowledge	1.10
Q9_NewIntegrationLimits	0.5 and 6 of 0.25	.70	F	.66	.60	.43	Comprehension	2.1
Q11a_Jacobian	1	.78	F	.68	.38	.41	Knowledge	n/a
Q12_NewIntegrand	4 of 0.5 and 1 of 1	.52	F	.65	.52	.50	Comprehension	2.1; 2.4
Q6_OptimumLagrange	6	.64	W	.64	.44	.49	Knowledge	1.12
Q10_TripleIntCylinder	6	.70	W	.63	.80	.58	Analysis	2.1 ; 2.2 ; 2.3 ; 2.4

Note. Questions marked with [a] means that this section is causing the item to be in question. When mode has been starred, it means that something from the distractor analysis has revealed something.
Original Cronbach is .70

Analysis of Results from the 2017 pilot.

The pilot in 2017 did not have an equivalent written component, so comparisons could not be made between writing the exam written, and then having it digitally assessed. Thus, looking at p-values, discrimination and where applicable, the chosen distractors was used for analysis about to what extent digital testing can be done.

Once again, sub-question 1 looked at the Blooms Taxonomy of the different items, as well as the course goals. The questions covered an acceptable range of course goals and contained various digital testing questions at Bloom's Application level. Sub-question 2 investigated the difficulty of items using p-values. The overall distribution of all three parts of the exams were similar, except in the open answer questions that two peaks in the 100% and 0% bars. This is due to the highly discriminating question 10. However, the open, final answer and MCQ parts of the exams did not correlate well. The highest correlation was between the final answer and written questions of 0.63. Question 2 showed a particularly low p-value and was marked as worrisome. Sub-question 3 investigated the discrimination of items, and all items have a surprisingly good two-group discrimination, showing a good ability to separate students in terms of ability. Only question 1 and 8 fall below the .30 line, and this is due to these items being easy. This is not problematic, due to the good psychological effect it has on students to have some easy items. However, investigating the patterns of responses of items, questions 2, 3 and 4 showed interesting patterns. Question 2 has one of the distractions answered more than the correct answer, and the RAR value shows a positive value just above 0, meaning that this option was chosen by an equal amount of weak and strong students. The correct answer is 25 and this distractor is 52, but with the great percentage students choosing the distractor, it is not credited to dyslexia. This question is however questionable in terms of its validity. Students had to read a lot of information with the use of a table and then read the wrong values for the chain rule. However, this distractor is a common misconception for students and the two-group discrimination did score above .30, showing that the very top students did still manage to do better on the item than weak students. Re-using this question without minor changes should however be reconsidered by test setters in a future pilot and compare results. Question 3 has an interesting pattern where many of the options seemed to be appealing. These options are filled with feasible common errors and small calculation mistakes. For example, the most chosen distractors is a mistake of not dividing my the unit vector. This question has a low CITC, meaning that the responses do not align well with the responses to the rest of the exam. However, the two-group discrimination is high, and this also comes through in the analysis of the distractors, where none of the commonly chosen distractors show a positive correlation. This is thus a good item, but some may be concerned about including responses that occur when a student only forgets a minus sign. This should be kept in mind when re-using this question to see if that aligns with the goals of the test. Question 4 also has a low CITC, but a very high two group discrimination. What is interesting is that the question asks

for an equation, but the most chosen distractor is not an equation at all. Only if this expression were to have " = 0 " at the end, it would be correct. Thus, it could be that students that got the right answer, but did read carefully and did not check their work, got this answer wrong. Again, validity might be compromised, as students have done all the correct calculations, but made the wrong choice. The question did well at discriminating, but the construct it is measuring is also "reading well". This detail should be reconsidered when re-using the item.

Whilst much of this discussion was about the specific items in the 2017 exam, it is also worth noting that only a few of all the items were discussed as "troublesome", and even these were very good items in terms of discriminating. Questions that were not discriminating well, were easy items. Interesting patters in multiple choice questions can be learned from to set better multiple-choice questions. None of the final answer questions were troublesome, as changes were made to assessing smaller mark increments were made. Unfortunately, none of the digital items went higher than application, but application goes above the hypothesis that digital questions will be simple recall. It was noted by the content expert that if levels of higher cognitive thinking wanted to be achieved in some other parts of the exam, instead of linear transformations, non-linear transformations could be used, as these are more cognitively complex.

Analysis of Evaluation Results from the 2016 and 2017 pilot.

Evaluation results were similar from students between 2016 and 2107, except that the digital testing construct went from worrisome to reasonable. Contrasting the responses of students and lecturers in similar areas, can be seen in table 25.

Students and staff agree that losing all your marks on just one mistake is troublesome and could be regarded as shallow assessment for Mathematics. Concerns are raised about it is digital testing really tests what you want to test in Mathematics, concerning argumentation. There is agreement also that digital testing is very good diagnostic exams, and that there should in fact be more of it. An interesting difference in opinion is that students are wanting 50% or less to be digital, whilst some staff believe that in the near future up to 80% can be digitally tested.

Table 24

Evaluation Responses Regarding Digital Testing from Staff and Students

Students	Teachers
Before the exam	
Good for diagnostic (+)	More Diagnostic should be done (+)
	Time writing the items (-)
	Difficulty in writing the items (-)
	Calculus has an item bank (+)
	Pilot new items
	More sharing between Universities (+)
The exam	
Stressful (-)	Checking process and answer should be mandatory learning goal
Unfair (-)	If the test does not test what you want to test, then advantage not worth it(-)
Method more important than answer (-)	
One error loses all marks (-)	Final Answer is shallow assessment (-)
	Punishing small mistakes is hard to justify (-)
	Faults can arise in many ways (-)
More of the questions should be written (-)	More questions can be digital (+)
	A way to justify answer is useful
Chromebooks (-)	
Fraud concerns (-)	
After the exam	
Should do human checking also (-)	Statistics are useful (+)
	Only saves time with big groups (-)

Note. the (-) symbols indicated a negative comment, a (+) symbol a positive comments and nothing indicated a neutral comment.

6. Discussion and Conclusion

This thesis investigated the question of "*To what extent can digital testing be included in first year calculus summative exams, for Engineering students?".* In order to answer this research question, four sub-questions were investigated:

1. Which differences are there in digital or written questions in meeting course goals at different cognitive levels?

2. To what extent can digital testing questions create the expected distribution of students according to their mathematical ability?

3. How is overall and item-wise discrimination effected by digital testing?

4. What is the current state of acceptance of digital testing calculus amongst staff and first year engineering students?

Below, the discussion of each sub-question.

Sub-question one: Which differences are there in digital or written questions in meeting course goals at different cognitive levels?

Both pilots meet an acceptable range of course goals and have attained a range of Bloom's Taxonomy beyond Knowledge and Understanding. In terms of the Educational Targets, it was expected that the digital questions would be able to meet these and in both exams 67% of the goals were covered. In terms of Bloom's Taxonomy, the hypothesis was that the digital questions would only be able to meet the Knowledge and Understanding levels – known as "basic skills". In 2017 multiple choice question and in 2016 final answer questions met a level higher than Knowledge and Comprehension, which is contrary to the hypothesis. An unexpected result in that one of the written questions in the 2017 pilot is a knowledge question. There is thus no notable difference in which level of Bloom's Taxonomy or course goal a written or digital question can meet.

Sub-question two: To what extent can digital testing questions create the expected distribution of students according to their mathematical ability?

This second sub research question wanted to investigate the ability of digital testing questions in creating the expected spread in terms of mathematical difficulty. Contrary to hypothesis, in the digital testing questions were able to create an appropriate spread, but as seen in the 2016 pilot with a higher average p-value, shifting the curve to the right. Upon analysis of items it could be seen that digital testing items were not only "easy" but could also achieved p-values below .3 and that of between .3 and .5. In fact, Multiple choice question were able to reach each category of difficulty from very difficult to very easy. The very difficult questions could be regarded as worrisome items, but with small changes

these items could be greatly improved in terms of validity, as discussed under the "Analysis of Results" section.

Sub-question three: How is overall and item-wise discrimination effected by digital testing?

The third sub-question investigated discrimination of digital items. This was done by looking at groups as well as item-specific discrimination. The hypothesis was that the middle academic group would disadvantage from digital testing. The hypothesis was wrong, as no academic group was specifically advantaged or disadvantaged from digital testing.

Through looking at the change in performance of groups when paper exams are assessed digitally, major findings of this sub-question include that in this 2016 group of 55 students, 14 students would pass in the written exam, but would fail in the digital exam. That is 25.5% of the group. A quarter of students failing where they traditionally would not. Likewise, 37 of the 55 students lose 0.5 marks or more in digital testing, which is 67%. This means two-thirds of students would drop in a mark in the digital exam compared to the written exam. Another finding is that many of the digital questions were very good at discriminating between weak and strong students. Items that did not discriminate well, were usually easy on purpose, as discussed in the section "Analysis of Results" section.

Analysis of the discrimination of distractors showed that having distractors with small mistakes such as forgetting a minus or forgetting to divide as troublesome, as even good students fall for these distractors, effecting the item's ability to discriminate.

Sub-question four: What is the current state of acceptance of digital testing calculus amongst staff and first year engineering students?

Digital acceptance among staff and students were investigated. Among students a quantitative questionnaire and a open answer qualitative answer was used and among staff a focus group was used. It was expected that the digital testing acceptance would be low.

In this study, the digital acceptance amongst students show that in the 2016 was low, falling below the 2.5 mark. This below the 2.5 mark can also be seen in the means of many of the questions, such as Questions 4, 5 6 8 and 9. This is accompanied by the negative comments that students had regarding the dislike, stress and unfairness of a digital exam.

In the 2017 pilot, the digital acceptance among students is considered reasonable, as it falls above the 2.5 mark, but still below the 3.5 mark. The change from 2016 is significant and could be due to the change from a 100% digital exam to a 66% digital exam, or the reworking of the digital items from 2016 to 2017. The comments in the open answers also include that of losing all marks for one mistake, digital testing being unfair and comments on how there should be more written questions. Upon analysis of the evaluation questions asking about the ratio of digital to written, it can be seen that 78.5% of the students that filled in the questions would like 50% or less of the exam to be digital, or that 45% of the

students that filled in the question wanted the digital testing component to be 25% or less, whilst only 9% wanted no digital testing at all. It could be assumed that these students have in mind that these 50% or 25% questions are the more "basic questions" – as expected in the hypothesis. Thus, whilst the qualitative answer might indicate a reasonable acceptance towards digital testing, the open answer and preferred ratio indicate that students would like the digital testing questions to consist less than the current 66% in the pilot, but that they are not against some parts of the exam being digital.

The focus group was organised by the lecturer and six of the fifteen lecturers invited came. Contrary to the hypothesis, the lecturers were overall positive of digital testing during the focus group and mention was made that up to 80% of the questions testing calculus for first year engineering students could be tested digitally. However, remarks made by lecturers indicated that they prefer multiple choice questions over final answer questions. There was a great awareness of caution that needs to be taken in writing these questions and that this costs a lot of time. Knowing how students will react to a new digital testing question was of concern, and pilots to try these out were suggested. Solutions regarding these obstacles of time and quality were that an item bank along with statistics of how items performed should be kept and shared among staff members of different universities. Other important remarks include that students should be taught to check their answer, and this should be a mandatory learning goal. Positive comments shared by both staff and students were that diagnostic testing for Mathematics is useful and should be done more often.

Limitations of Research

Sub-question one: Cognitive Levels

The coding of these questions is somewhat subjective. Much criticism is against the use of Bloom's Taxonomy for classifying Mathematics questions, as much of it depends on what the learner has seen or has practiced. A complex analysis question could become a trivial knowledge question merely by being taught or seen before in a homework exercise. Thus knowledge about how students experience questions based on their previous experience is as best, a calculated guess. To minimise this, a more objective way of classifying the questions were attempted in writing the Bloom's Taxonomy classification, with specific examples and what is expected of the learner to do. The Educational Targets also have limitations as during the coding the content expert expressed that whilst they might code a question with a certain educational code, this was only because it meets one part of the sentence. Thus at times the questions would seem to meet many goals, but only a part of each – whilst another question would meet only one educational goal, but it means all parts of that goal. Thus, by doing a simple count – is not enough. Also, only using one content expert for the scoring of the Bloom's Taxonomy is thus not advised and more than one should be used. It was hoped that the course goals would also be classified due to their level in Bloom's Taxonomy, but as the course goals were not written with Bloom's Taxonomy in mind, being limited to the verb "Apply", this was not possible in this research to

make a link between the cognitive level of the question and the cognitive level of the course goal. Rewriting the course goals with Bloom's Taxonomy in mind could have the potential for reaching a wider range of cognitive competencies in the goals of the course.

Sub-question two: Difficulty spread

Classical test theory is sensitive to the number of students that write the item. Thus, it is more representative of the group, than of the item. In the 2016 pilot, the p-value analysis was limited to one class of 55 students, which has implementations upon the external validity of the results.

Sub-question three: Discrimination

In the 2016 pilot, the majority (91.7%) of the questions were final answer questions, limiting what can be said about Multiple-Choice questions in terms of what happens to discrimination when digital assessment is used. The focus of much literature when it comes to digital testing is on MCQ, so this is unfortunate. However, this can be a contribution to literature for the very same reason.

Upon further investigation it might appear that the middle group in final answer questions becomes the most disadvantaged in certain items, but this could be due to a ceiling effect, and the results cannot be generalised.

Sub-question four: Digital Acceptance

Possible limitations in the digital acceptance variable would be that the survey was not designed beforehand to measure this construct. Thus at least three variables could be chosen for this construct. However, this construct did align well with what students thought in the open answers. Only two of the questions used for the digital testing construct in the 2016 and 2017 pilot were the same. This might have consequences of not being comparable.

For the focus group, three digital testing project members were present. This could have made the overall atmosphere to be in favour in digital testing. Attendance to the focus group was voluntary, meaning that those interested in digital testing were the most enthusiastic in attending. One lecturer could not come as they do not teach on a Friday but had more negative opinions that they emailed. The planning of a focus group should be done in future on a Wednesday or Thursday when most maths staff at the University.

Overall Conclusion

This thesis investigated the question of "*To what extent can digital testing be included in first year calculus summative exams, for Engineering students?"* It was found that digital testing questions can assess higher cognitive levels according to Bloom's Taxonomy. Also, digital testing items were not just "easy" or measuring "a limited set skill". The p-values showed that there are difficult items possible, even though these items were of a lower cognitive level in Bloom's Taxonomy. One question that the 2016 pilot had was a good digital testing item with a

desirable p-value around .60, and measured the cognitive level of synthesis. Whilst there are many concerns about the time to create good items, this item was taken out of an item-bank. However, a content expert is still needed to know what to look for, and this could take time. The 2016 pilot revealed that the possible strength of this item was that it was a final answer questions with 1 mark or 0.5 mark increments. Final Answer questions with 2 or 3 marks seems to be beyond the extend of what should be measured with final answer at this present time. In the 2017 pilot, more analysis was done in terms of what is possible with multiple choice. Here items were also kept to two or three marks, and with simple calculations. Whilst the items showed some interesting pattern responses indicating that items distractors with small calculations mistakes also "distract" strong academic students. However, omitting these answers in the distractors are also problematic that would make the question easier. A new pilot is needed with these same questions, making changes due to mistakes that occur due to bad reading and trying final answer format instead of MCQ. Despite these response patterns, these questions still performed well in terms of their discriminating power, showing that what they were measuring is indeed within the extent of what is possible in terms of digital testing.

With the increase of digital acceptance from 2016 to 2017, from worrisome to reasonable, it shows that the changes made from the 2016 to the 2017 pilot, is going in the right direction. However, students are still unsatisfied with the number of items that are being digitally tested, indicating that a 50% digital exam would be more acceptable than the 66% digital of 2017. Some lecturers indicated that 80% of exam questions can be set digitally, yet with the time needed to set digital testing questions and with the low acceptance amongst students, it might be wise to not go to 80% digital and continue with pilots with a lower percentage such as 50%. As mentioned in the focus group, technology is also changing, with exciting possibilities of AI and redoing questions for new marks. As these become available soon, this will widen the scope of to what extent can be tested, fostering new exciting possibilities in both research and education.

7. Reference list

Azevedo, J., Oliveira, E., & Beites, P. (2017). How Do Mathematics Teachers in Higher Education Look at E-assessment with Multiple-Choice Questions. *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017)*, *2,* 137-145. doi: 10.5220/0006324801370145

Backhoff, E., Larrazolo, N., & Rosas, M. (2000). The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electrónica de Investigación Educativa, 2*(1). Retrieved from: http://redie.uabc.mx/vol2no1/contents-backhoff.htm

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objetives: the classification of educational goals: handbook I: cognitive domain* (No. 373.19 C734t). New York, US: D. Mckay.

Burton, S., Sudweeks, R., Merrill, P., & Wood, B. (1991). How to prepare better multiple-choice test items: guidelines for university faculty, *Brigham Young University Testing Services and The Department of Instructional Science.* Retrieved from: http://testing.byu.edu/info/handbooks/betteritems.pdf.

Chalies, N., Houston, K., & Stirling, D. (2004). Supporting Good Practice in Assessment in Mathematics, Statistics and Operational Research. *The Higher Education Academy.* Retrieved from: https://www.heacademy.ac.uk/system/files/supportinggoodpractice.pdf

DiBattista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-choice Items on Classroom Tests. *The Canadian Journal for the Scholarship of Teaching and Learning*. *2011, 2*(2). doi: http://dx.doi.org/10.5206/cjsotl-rcacea.2011.2.4

Ding, L., Chabay, R., Sherwood, B., Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Physics Education Research,2*, 1-7. doi:10.1103/PhysRevSTPER.2.010105

Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement. Retrieved from: https://ebookppsunp.files.wordpress.com/2016/06/robert_l-ebel_david_a-_frisbie_essentials_of_edbookfi-org.pdf

Evertse, J. (2014). Digital testing: Opportunities for higher education. Insight into the effects of digital testing and experience gained from the testing and test-driven learning programma, 2010-2014. Editorial: Daphne Riksen and SURF. Retrieved from: https://www.surf.nl/binaries/content/assets/surf/en/2014/impact-

assesment-report-digital-testing-opportunities-for-higher-education-2014.pdf

Gainsburg, J. (2007). The mathematical disposition of structural engineers. *Journal for Research in Mathematics Education*, *38*(5), 477–506. doi: 10.2307/30034962

Gibbs, G., & Simpson, C. (2005). Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education, 1*, 3-31. Retrieved from: http://eprints.glos.ac.uk/id/eprint/3609

Radmehr, F., & Drake, M. (2017). Revised Bloom's taxonomy and integral calculus: unpacking the knowledge dimension. *International Journal of Mathematical Education in Science and Technology*, *48*(8), 1206-1224. doi: 10.1080/0020739X.2017.1321796

Haladyna, T. M., Downing S. M., & Rodriguez M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied measurement in Education, 15*(3)*, 309-334. doi: 10.1207/S15324818AME1503_5*

Impara, J. C., and Foster, D. (2006). Item and Test Development Strategies to Minimise Test Fraud. In Handbook of Test Development: Chapter 5 (pp. 91 – 114). doi: 10.4324/9780203874776.ch5

Kastner, M., & Stangl, B. (2011). Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter*. Procedia Social and Behavioral Sciences, 12,* 263-273. doi:10.1016/j.sbspro.2011.02.035

Karaali, G. (2011). An Evaluative Calculus Project: Applying Bloom's Taxonomy to the Calculus Classroom. *PRIMUS*, *21*(8), 719-731. doi: 10.1080/10511971003663971

Kent, P., & Noss, R. (2002). The mathematical components of engineering expertise: the relationship between doing and understanding mathematics. *Paper submitted to the Institution of Electrical Engineers Annual Symposium on Engineering Education, London.* Retrieved from: http://www.oemg.ac.at/FH/Klagenfurt2005/Kent-Noss-EE2002-preprint.pdf

Lee, S., Harrison, M. C., & Robinson, C. L. (2008) Identifying what makes a good question in a mechanics diagnostic test. *International Journal of Mechanical Engineering Education*, *36*(8)*,* 256-265. doi: 10.7227/IJMEE.36.3.9

Myers, C. T. (1955). The relationship between item difficulty and test validity and reliability. *ETS Research Bulletin Series*, *1955*(2), i-7. doi: 10.1002/j.2333-8504.1955.tb00257.x

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education, 31*(1), 53-64. doi: 10.1080/03098770601167922

Odukoya, J., Adekeye, O., Igbinoba, A., & Afolabi, A., (2018). Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Quality & Quantity: International Journal of Methodology*, *52*(3), 983-997. doi: 10.1007/s11135-017-0499-2

Paterson, J. S. (2002). *Linking on-line Assessment in Mathematics to Cognitive Skills*. Paper presented at the Proceedings of the 6th CAA Conference, Loughborough: Loughborough University. Abstract retrieved from https://dspace.lboro.ac.uk/2134/1887

Robinson, C. L., Harrison M.C., & Lee, S. (2005) *Engineering Subject Centre Report: Responding to the Changes in the Teaching and Learning of Mechanics in Schools.* Higher Education Academy-Engineering Subject Centre. Retrieved from: https://www.heacademy.ac.uk/system/files/responding-to-changes-teaching-learning.pdf

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1), 1-11. doi: http://dx.doi.org/10.1080/2331186X.2017.1301013

Sangwin C. J., & Köcher, N. (2016). Automation of mathematics examinations. *Computers and Education*, 94, 215 -227. doi: http://dx.doi.org/10.1016/j.compedu.2015.11.014

Shorser, L. (1999). Blooms Taxonomy Interpreted for Mathematics. Retrieved from: http://www.math.toronto.edu/writing/BloomsTaxonomy.pdf

Torres, C., Lopes, A. P., Babo, L., Azevedo, J. (2009). Developing Multiple-Choice Questions in mathematics*.* Paper presented at the *Proceedings of ICERI 2009 - International Conference of Education, Research and Innovation.*

van der Wal, N. J., Bakker, A., & Drijvers, P. (2017). Which Techno-mathematical Literacies Are Essential for Future Engineers?. *International Journal of Science and Mathematics Education*, *15*(1), 87-104.

Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *The National Medical Journal of India*, *21*(3), 130-133. Retrieved from: https://www.ncbi.nlm.nih.gov/pubmed/19004145

Appendix 1: Exam Questions, Pilot 2016

Deliberately removed, please contact the author for the any questions.

Appendix 2: Evaluation Questions, Pilot 2016

Questions asked at the end of the pilot test done in June 2016. Questions 2 until 11 were answered on a five point Likert scale, whilst question 12 was open ended. The original Likert scale was 1 = fully agree; 2 = agree ; 3 = neutral ; 4 = disagree ; 5 = totally disagree. The scale was flipped, as below, to align with the pilot of 2017, and questions 2 to 11 recoded before analysis to align with the questionnaire below.

1 How much additional time did you spend in entering your answers of the math X exam in MyLabsPlus?

1 = 0 - 5 min. 2 = 5 - 10 min. 3 = 10 - 15 min. 4 = 15 - 20 min. 5 = >20 min.

2 I have a lot of experience in entering answers in MyLabsPlus.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

3 I did not have problems entering my answers in MyLabsPlus.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

4 It would be fair if my final answers of the exam will be graded, and not my handwritten calculations and argumentations.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

5 It would be an advantage for me if only my final answers of the exam will be graded, and not all my handwritten calculations and argumentations.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

6 I would work more accurately if only my final answers would be graded, and not my handwritten calculations and argumentations.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

7 If the system would check each of my answers directly after each complex question and would give me another try when my answer was wrong (e.g. due to a typing error or an error in calcultating) but for less points, this

would be a fair way to get a grade that is based on final answers only (and not on any calulations or arguments).

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

8 I believe that a digital Math exam with MyLabsPlus using the Respondus Lock Down Browser is a good way to test my knowledge and skills.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

9 I believe that a digital Math exam on my own laptop/device is a fair way to test my math knowledge and skills.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

10 For me it is not a problem to bring my own device for doing a Math exam.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

11 I believe that a digital Math exam on my own laptop/device is fraudproof.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

12 Suggestions/comments:

Appendix 3: Exam Questions, Pilot 2017

Deliberately removed, please contact the author for the any questions.

Appendix 4: Evaluation Questions for Pilot 2017

Section A: Questions asked at the end of the pilot test done in June 2017 to **all students** completing the pilot test on paper.

1 I believe a hybrid Math exam with both short answer questions (e.g. multiple choice) and open questions with written solutions (incl. calculations) is a good way to test my knowledge and skills.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

2 I believe a hybrid Math exam with both short answer questions (e.g. multiple choice) and open questions with written solutions (incl. calculations) is a fair way to test my knowledge and skills.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

3 I believe that the short answer questions in this exam were easier than the open questions with written solutions (incl. calculations)

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

4 The ratio between short answer questions (2/3th of all points) and open questions with written solutions incl. calculations (1/3th of all points) in this exam is right.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

5 Which ratio between short answer (S) and written solution (W) questions do you prefer?

1 = 100% - 0% 2 = 75% - 25% 3 = 50% - 50% 4 = 25% - 75% 5 = 0% - 100%

6 I would have done better on a traditional paper exam than the hybrid exam I took today.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

7 I believe that the selfstudies, tutorials and the diagnostic and sample tests were a good preparation for this hybrid exam.

1 = yes 2 = no

Section B: Questions asked at the end of the pilot test done in June 2017 to **Electrical Engineering** students after the hybrid exam done in MyLabsPlus and on paper.

1 I believe a hybrid Math exam with both short answer questions (e.g. multiple choice) and open questions with written solutions (incl. calculations) is a good way to test my knowledge and skills.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

2 I believe a hybrid Math exam with both short answer questions (e.g. multiple choice) and open questions with written solutions (incl. calculations) is a fair way to test my knowledge and skills.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

3 The ratio between short answer questions (2/3th of all points) and open questions with written solutions incl. calculations (1/3th of all points) in this exam is right.

1 = yes 2 = no

4 *If no, please choose one of the following ratios of "Short Answer – Written":*

1 = 100% - 0% 2 = 75% - 25% 3 = 50% - 50% 4 = 25% - 75% 5 = 0% - 100%

5 I believe that the short answer questions in this exam were easier than the open questions with written solutions (incl. calculations)

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

6 I would have done better on a traditional paper exam than the hybrid exam I took today.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

7 I believe that the selfstudies, tutorials and the diagnostic and sample tests were a good preparation for this hybrid exam.

1 = yes 2 = no

8 I did not have problems entering my answers in MyLabsPlus.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

9 The WIFI strength during the test was sufficient.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

10 There were no technical issues during the test.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

11 Did you encounter any hardware issues during the test?

1 = no problems 2 = mousepad 3 = sound 4 = keyboard

5 = external mouse 6 = screen 7 = battery 8 = other

12 It would be a better way to get a grade that is based on final answers only than the current exam, if MyLabsPlus would check my answers directly after each question and would give me another try when my answer was wrong, but for less points.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

13 I prefer MyLabsPlus to show my score directly after answering each short answer question.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

14 I believe that a digital exam in MyLabsPlus on a Chromebook is fraud proof.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

15 I prefer a digital exam to a handwritten exam.

1 = totally disagree 2 = disagree 3 = neutral 4 = agree 5 = totally agree

16 Other remarks or suggestions for improvement:

Appendix 5: Invitation to lecturers for focus group.

Email Subject: Focus group regarding digital testing in mathematics exams

Dear …..

I would like to invite you to a free lunch, where you will have the opportunity to express and discuss with other lecturers your thoughts digital testing. This will include the advantages, concerns and question item types in digital summative calculus exams.

Your opinion will be voice recorded but remains anonymous when included in my master's thesis, which has a research question: "*To what extent can engineering students have their first-year calculus course tested in a summative digital exam?*". I believe that your opinion and expertise is invaluable in answering such a question.

During the focus group, I only ask that you supply me with the number of years of teaching experience that you have and any experience that you have with digital testing, so that I can validate my data. Your participation is voluntary, and you can withdraw from the study at any time. A summary of ideas and points raised will be documented after the focus group and sent back to all participants.

If you can come, I would very much appreciate your attendance on Wednesday the 17th of October 2018 at [excluded due to privacy] from 12:00 until 13:30. Please let me know if you can attend, along with any dietary requirements, so that I can make the necessary arrangements.

If you have any further questions, or would like to offer an opinion on digital testing but cannot come to the focus group, please contact me on:
a.j.lochner@student.utwente.nl
Kind regards,
Alisa Lochner
Educational Science and Technology master's student

Appendix 6: Focus Group Planning and Brainstorming

Math Digital Testing Focus Group 2018

1 . For the purpose
At the University of Twente, there has been since 2015 a project group investigating the possibilities regarding digital testing in undergraduate Mathematics. The initial research question that the research group investigated was "To what extent can undergraduate Mathematics be tested using MyLabsPlus". The data gathered in these pilots is being used in the master's thesis of Alisa Lochner to answer a similar question "To what extent can digital testing be included in first year calculus summative exams, for engineering students?". To answer this research question, digital testing pilot items have been analysed. However, the opinion and experiences of students and staff is also important in answering the question. During the pilot tests students had an opportunity to answer and express their opinions. In terms of staff, informal interviews have been done and literature shows that teachers of Mathematics are not welcoming towards using digital testing. It would like to be discovered, through this focus group, if this is also the case here and now at the University of Twente, and what the reason for this would be.

2. Potential participants
Ideally, we would like about 6 to 10 people.
The list of people we could ask are (*__bold__ are coming):

[Excluded due to privacy of participants]

3. Time, Date and Location:

Time:	Date:	Location:
12:15 until 13:30.	Friday 19 October 2018	[Excluded due to privacy]

4. Core questions/goals
Since there is a limited amount of time, the main questions should be limited to three. All questions should be open questions.
The first question should be an introductory question, about issues around digital testing, before focusing on certain aspects. Then the following two questions can ask the main purpose of the day.

Key Questions
To facilitate the discussion, a short talk will be given by Alisa about digital testing. The print-outs of the 2017 pilot will be available as question types.

Introductory Question: What is your first impression regarding the advantages of these question types for Mathematics, as in the 2017 pilot?

Key Question 1: Would you use these question types in an exam that you were setting? If not, why not? (So in a sense, disadvantages)

Key Question 2 (choice):
"What possibilities/question types would you like to see in digital testing?"

"Hypothetically speaking: Say that digital testing becomes the norm at the University, what kind of support would you as a lecturer would like to receive?"

Probing questions

"Could you tell me a bit more about that?";
"I'm not quite sure what you mean….?";
"Could you explain a bit more?"
"How does that work in practice?"
"Can you give us an example?"

Question to avoid

Be aware of:
leading questions e.g. "They think this, how about you?"
"emotionally charged questions"
"double barrelled" questions.
Closed questions.

Questions Brainstorming

What possibilities/question types in digital testing would you like to see?
What kind of support regarding summative digital testing would you like to receive?
Do you feel that summative digital testing with Mathematics is the future?
What do you think is the next step in digital testing?

5. Collecting Information

A voice recorder will be place in the middle of the room for recording purposes. These transcripts and summaries will be emailed back to the participants for their further input.

6. Permission that needs to be gathered by respondents

A permission slip will be handed out to participants. This includes the purpose of the focus group, how information will be treated confidentially as well as how the data gathered will be examined and processed. It will be made clear to participants that they can agree to not having their data used at any time.

7. Materials needed for the focus group

Room requirements:
A projector plus a screen.
Seating that is for about 10 people.
Materials brought to the room:
Print-outs of 2017 pilot exam.
Refreshments (Brought by 12:15).
Permission papers.
Name tags
Tables either in a circle or a U-shape.

Appendix 7: Focus Group Summary

Section A: 19 October 2019
This focus group was conducted with six lecturers. Their experience in teaching ranges from 2.5 to 38 years (17.4 years average). Not all experience is at the University of Twente, but also at other Universities. Only one of these lecturers has never had any of their courses tested digitally. Most lecturers had some experience with some of their courses being tested digitally, with Multiple Choice being mentioned the most. Experience ranged from making items in Maple TA, having half a course tested with Multiple Choice for 6 years and organising the digital platform for a university. Two others besides the researcher were present that mainly posed questions: an educational advisor for the mathematics faculty and an associate professor from ELAN (Department of Teacher Development) that is and has been involved with digital testing projects and assisting lecturers that want to adopt digital testing in their courses. Below a summary of some of the main points discussed.

Advantages
It saves a lot of time during assessment, especially with large numbers. It is a problem of scale. With small groups this advantage is not realised.

Statistics gathered about items can give more information about which kinds of items students find easier and more difficult – not only aiding item selection, but also item position in an exam, where ideally the easiest items are at the start and the more difficult items are at the end.

Digital testing gives access to a lot more statistics. More advantage can be taken of this during the course, using diagnostic testing. One the advantages of this is getting to know your group better, and you can possibly adapt the exam to the group.

Digital testing is not lenient towards sloppy work. If you are dealing with engineers whom later make mistakes and endanger lives, e.g. parking lots collapse, this could be used to an advantage if/when used correctly.

Digital testing is good for basic questions, and it is expected that 80% of the questions for summative exams for engineers can be handled digitally.

Multiple Choice Questions (MCQ) gives more information than final answer questions to the teacher about the misunderstandings that students have. MCQ has the advantageous possibility of awarding "half a mark" to certain distractors when chosen, whilst with final answer this is much more complex to do.

Disadvantages/Concerns
More time is spent on the writing of the items, as writing quality items/distractors needs time and creativity.

Multiple choice items are limited in what can be asked using them, for example – you cannot ask a "solve this equation" question and then give four options – resubstituting will occur and reverse engineering and this is not one of the educational goals.

Students get punished for small mistakes in digital testing and that can be hard to justify to both staff and students.

Final answer questions have various concerns – one of which that with current systems it seems to be an all or nothing approach, whilst multiple choice offers a type of "safety net".

The use of final answer appears to be very shallow assessment if it is all or nothing due to one small mistake. Whilst some assessment is open to that, like basic skills, but then setting those questions are really hard.

Expectation management – warning students about how digital assessments is strict – and careful item construction can help but not completely prevent all problems that come with digital testing.

Digital Testing might save time, but if it still does not test what you want to test – it can save as much time as you want, but if it does not test what you want it to test, the feeling is then "so what if it saves time".

Neutral comments
Small mistakes are punished in digital testing, but it is not the philosophy behind the it and that needs to be made clear.

Don't know how the students will react to new digital items – need a pilot to determine it.

For calculus there are a lot of items in the item bank available.

Checking your process and answer should be a mandatory learning goal. There needs to be more focus in preparing students for the digital exams and how they lose can check to not
loose points. A sample exam helps, but more needs to also be done.
Information gained in digital testing is not fed back to the group, as digital testing is mainly done at the end of a course. More formative assessments should be done where 10% of the questions are used, which can inform as to what kind of group are you dealing with.

To take into consideration when pursuing digital testing

Sharing of resources is very important and valuable. Helping with the setting of exams. Sharing within 4TU is a good start.

An item back of 10 -100 examples of good examples is a good form of support that teachers would need to know what works. These questions should come with statistics about how students perform on certain types of questions – in order to help the validity, quality and the speed of setting exams

About 80% of exam items can be based on an item bank, whilst the other 20% needs to be original new multi-step items that are more creative. The type of questions in such an item bank should also be carefully selected – as once a student sees a problem, they know how to do it next time and then it becomes a trivial problem, for example finding a second-degree polynomial at a certain point. The 20% more creative problems should not exactly test the learning goals but be multi-step procedure questions that cause you to see how students apply the learning goals. This is because if you have one goal per question, then depending on what the question looks like – students know what to do. With these more multi-step questions the answers diverse and it is impossible to give a fair grade just based on final answer. It might be discouraging to put these questions even in a written exam, because it is already known that students will perform badly. However, if students expect these questions – they will prepare for them. It all comes down to "What do you want to assess".

In setting digital exams, teachers are semi-confident in the setting of multiple-choice questions but would require assistance working with the system and implementing their items in a digital environment. Other comments were that the management of the item bank would require a programmer more than a maths teacher.

Simplicity is key in exams. Questions with too much detail and information and tables cost a lot of reading time and may decrease motivation- especially for those with dyslexia. Simplicity does not mean that the exercises is easy. An example of a complicated question is question 2 from the 2017 pilot. However, if future classes also do badly on such a question – you can tell if it is too much information – or if students just were not used to answering a question in this way.

Considerations for future research

The alternatives in a multiple-choice question – perhaps an idea for one of the alternatives be a halfway step, but only worth half the marks of the full answer.
With more sophisticated systems, it is possible to code your own evaluation system where you can, for example, program half marks for common mistakes

and full marks for the right answer. Might even be possible to see what students are typing in.

Future research could think of employing machine learning to learn from the past old exams. This could help to learn from the past in making distractors and give marks to final answer possibilities, or on how to improve future designs of exams. Currently industry is quite active in looking at applications in Artificial Intelligence and could be useful for application in assessment over 5 or so years.

Future research in human media interaction could also make use of language processing where 10 words or maximum of 3 lines can probably be assessed in three to four years. More than that, you can program into it five most common answers to look for first. At the moment just keywords are recognised and thus the argumentation between these keywords are really missing. Seeing that even human emotion can be detected from spoken language, most surely we will get to a point where the argumentation between keywords are also understood.

Could be interesting to see how students would perform on questions where they have to select no answer is correct, or they could select more than one answer to be correct. This prevents students from just "searching for the right answer" among these options in MCQ. Students would be given both positive and negative marks based on their choices. Negative marking is not favoured by students, so this would have to be done carefully.

Upon getting a question wrong, students get another change or an intermediate step for less points and/or upon getting a question right, students must justify their answer. The latter was commented as a nice model, but not necessary. The former suggestion concerns were raised that you are making weaker students take more time on the exam – and then the exam is harder for the students that already find it hard. A suggestion to counter this was that the maximum time for such an exam should be the time taken for the longest pathway taken. Another concern is that students already hesitate in such an exam as they are not used to it, incorporating something like this where they get something if they get something wrong that increases anxiety. This kind of feedback about what you got wrong is only something you usually get after an exam. Mention was made that if a move was made from just final answer exams to an exam that could do an intermediate or scaffolded question, this is a good option.

A box that students must check about how confident they are about their answer in a digital testing exam. The very existence of such a box might force students to check their work. In the conversation someone asked if the more confident the person is, the more points they get, but this is not necessary.

Section B: Email response by a lecturer that could not attend. The lecturer has more than 35 years experience in teaching.

I understand that increased student numbers seem to generate a need for digital testing, but in my opinion the students' knowledge that we can test this way is rather limited. At best one could add a few multiple-choice questions trying to check whether they understood certain concepts. So-called "do exercises" where the student is to compute a solution and input the result in a digital system are unsatisfactory because faults can arise in many ways. The problem becomes completely hopeless when it comes to "prove that ..." exercises.

I usually include some multiple-choice tests in exams for [….] students and this works well because I ask them to motivate their answer in one or two sentences. This does check knowledge quite well and is still easy to mark.

We also had some experiments with exercises where the students had just to write down the solution of a computational question in the math line. Students did not like that as far as I remember (but there was a questionnaire, so you should be able to retrieve that) because they were afraid of getting 0 points for simple computing errors. Neither did it save a lot of time since exam preparation was rather expensive:

With this kind of test, one has to make sure that writing over does not become too easy. So, there were 4 different versions of the exam to be prepared. Quite a few students seem to like mylabs+, so digital testing is good for training, but not for an exam, I would say. For an average exam I usually spend 10 min per student (at least after checking the first 10 or so). If we cannot afford this anymore, we should maybe hire more people?

Appendix 8: Mathematics X: Educational Targets

For easy of use, the Educational Targets have been put into numbered table. In all cases, the student is able to (especially w.r.t. functions of two or three variables):

Code	*Educational Target*
	Section 1: Work with partial derivatives and applications
1.1.	Apply the parametrization of a curve and the tangent vector
1.2.	Apply the chain rule (in several forms)
1.3.	Calculate a directional derivative, and apply its properties
1.4.	Calculate the gradient (vector)
1.5.	Apply the relations between gradient and level sets
1.6.	Calculate the tangent plane and normal line
1.7.	Apply a linearization (standard linear approximation)
1.8.	Estimate a change using differentials
1.9.	Calculate Taylor polynomials (first and second order, two variables)
1.10.	Apply the first and second derivative tests
1.11.	Calculate the absolute extreme values on closed bounded regions
1.12.	Apply the method of Lagrange multipliers
	Section 2: Define and evaluate double and triple integrals over bounded regions
2.1.	Sketch the region and find the limits of integration
2.2.	Calculate an iterated integral (by changing the order of integration)
2.3.	Define area, volume, mass or the average value as an integral
2.4.	Apply polar, cylindrical or spherical coordinate substitutions, or a given transformation
2.5.	Calculate centroid, (center of) mass and first moments

Appendix 9: Bloom's Taxonomy in Math

Taxonomy Table constructed from Shorser (1999) and Torres et al. (2009),

	Knowledge
Definition	Recalling memorized information: retention of terminology, facts, conventions, methodologies, structures, principles, etc.
Examples	Know common terms, specific facts, methods and procedures, basic concepts, principles. Questions include "State the definition", "State the theorem", or "Use the specified method." E.g., Take the derivative of the following rational function using quotient rule.
Keywords	Define, list, state, identify, label, match, select, describe, name, what, tabulate, when
	Comprehension
Definition	Grasping the meaning of material. Translation, extrapolation, interpretation of facts, making comparisons, etc.
Examples	They would be able to explain a graph, a calculation using a formula, or an equation. Understand a definition or a theorem and how they relate with other definitions or assumptions. Describe its first and/or second derivative. Questions ask the student to use definitions or methods to calculate something. E.g., Find the slope of the tangent line to the following function at a given point.
Keywords	Explain, predict, interpret, infer, summarize, convert, translate, give example, associate, estimate, extend, give counter-example, paraphrase Summarize, compare and contrast, estimate, discuss, etc.
	Application
Definition	The ability to use learned material in new and concrete situations Involves doing. Problem solving.
Examples	Students must be able to demonstrate that they can use concepts and theories in problem-solving. Apply concepts and principles to new situations. Construct graphs and charts, demonstrate the correct usage of a method or procedure. Implement a known strategy to solve an exercise in a new area. Questions which require the usage of more than one definition, theorem, and/or algorithm. E.g., Find the derivative of the following implicitly defined function. (This

	question could be used to test logarithmic differentiation as well, for instance)
Keywords	How would you show, modify, demonstrate, solve, or apply x to conditions y, calculate, operate, complete, classify, discover, relate, … Apply, calculate, complete, solve, modify,
	Analysis
Definition	Identifying parts, analysis of relationships between parts, recognition of the organizational principles involved. Students must be able to take a situation apart, diagnose its pieces, and decide for themselves what tools (graph, calculation, formula, …) to apply to solve the problem at hand. Ability to classify abstract structures and objects. Making inferences and supporting them with evidence, identification of patterns.
Examples	Recognize unspecified assumptions, recognizes logical misconceptions in reasoning, distinguish between facts and inferences, evaluate the relevancy of data, analyze the organizational structure of a work. Questions require the student to identify the appropriate theorem and use it to arrive at the given conclusion or classification. Alternatively, these questions can provide a scenario and ask the student to generate a certain type of conclusion. E.g., Let f(x) be a fourth-degree polynomial. How many roots can f(x) have? Explain.
Keywords	analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer, how does x affect or relate to y, why, how, diagram, differentiate, distinguish.
	Synthesis
Definition	In contrast to analysis (i.e., taking apart), at the synthesis level students put things back together. Integrate learning from different areas or solve problems by creative thinking. Relate knowledge from several areas. Generalize a theorem or prove a new theorem. Derivation of abstract relations, prediction, generalization, creation of new ideas
Examples	Integrate learning from different areas into a plan for solving a problem, formulate a new scheme for classifying objects. Include generalization from given facts, relating knowledge from several areas, predicting, and drawing conclusions.

	Questions are similar to Analysis questions, but the conclusion to be reached by the student is an algorithm for solving the given question. This also includes questions which ask the student to develop their own classification system E.g., optimization word problems where student generates the function to be difirerentiated.
Keywords	formulate, generalize, rewrite, combine, integrate, formulate, design, create, prepare, modify, rearrange, substitute, invent, what if, compose, construct,

Evaluation	
Definition	At the evaluation level, one is able "to judge the work of others at any level of learning with regard to its accuracy, completeness, logic, and contribution" (White, 2007, p. 161). The ability to judge the value of material. judgement of validity, usage of a set of criteria to make conclusions, discrimination
Examples	Compare and discriminate between ideas. Verify value of evidence. Questions are similar to Synthesis questions, except the student is required to make judgements about which information should be used. E.g., related rate word problem where student decides which formulae are to be used and which of the given numbers are constants or instantaneous values.
Keywords	Assess, rank, grade, test, measure, Appraise, Compare, Conclude, Contrast, Criticize, Describe, Discriminate, Explain, Justify, Interpret, Support

Appendix 10: Digital Testing Acceptance Construct Creation

The process of creating a construct a digital testing construct involved selecting appropriate questions. In order to make an appropriate selection, the reliability of items chosen were checked, along with the checking of factors and correlations.

Digital Testing Construct from the 2016 pilot

Table A

Reliability of All Items from 2016 Evaluation Questionnaire

Question	N	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q2_ExpMLP	55	.198	.620
Q3_InputMLP	54	.276	.606
Q4_DigiOnlyFair	55	.260	.609
Q5_DigiOnlyAdv	55	.237	.614
Q6_DigiOnlyNeat	55	-.008	.672
Q7_CheckAns	55	.367	.584
Q8_GoodWay	55	.658	.510
Q9_FairWay	55	.575	.536
Q10_OwnDevice	55	.278	.605
Q11_FraudProof	55	.204	.626
Total			.627

Note. Q1 was excluded due to being irrelevant and in minutes

Table B

Reliability of Digital Acceptance Construct Items in 2016 pilot before deleting Q5

Question	N	Corrected Item-Total Correlation	Cronbach's Alpha if Item deleted
Q4_DigiOnlyFair	55	.452	.697
Q5_DigiOnlyAdv	55	.466	.689
Q8_GoodWay	55	.468	.625
Q9_FairWay	55	.574	.621
DigiAcceptance2016	55		.722

Table C

Reliability of Digital Acceptance construct Items in 2016 pilot after deleting Q5

Question	N	Corrected Item-Total Correlation	Cronbach's Alpha if Item deleted
Q4_DigiOnlyFair	55	.378	.763
Q8_GoodWay	55	.583	.510
Q9_FairWay	55	.590	.505
DigiAcceptance2016	55		.697

Table D

Explorative Factor Analysis for "Digital Acceptance" in 2016 pilot

Measure	Factor 1	Factor 2	Factor 3	Factor 4
Q3_InputMLP	**.459**	.037	.149	.096
Q10_OwnDevice	.**827**	.041	-.085	.004
Q4_DigiOnlyFair	-.009	**.489**	.015	-.028
Q5_DigiOnlyAdv	.017	**.908**	-.331	.016
Q6_DigiOnlyNeat	-.339	.240	.122	-.084
Q8_GoodWay	.143	**.451**	.398	.332
Q9_FairWay	.356	.**483**	.210	.162
Q7_CheckAns	-.001	.018	**.601**	.085
Q11_FraudProof	.041	-.149	**.641**	-.056
Q2_ExpMLP	-.043	-.018	.003	**.810**
Q1_MLPTime	-.444	-.111	-.175	.131

Note. Factor loadings above .400 were considered high enough, and were bolded.

Table E

Pattern Matrix of the Digital Acceptance Factor from the 2016 exam.

	Digital Testing Acceptance
Q4_DigiOnlyFair	.515
Q5_DigiOnlyAdv	.546
Q8_GoodWay	.726
Q9_FairWay	.730

Table F

Pearson Corrolation for Digital Acceptance construct Items 2016

	Q4_DigiOnlyFair	Q5_DigiOnlyAdv	Q8_GoodWay	Q9_FairWay
Q4_DigiOnlyFair				
Q5_DigiOnlyAdv	.444**			
Q8_GoodWay	.314*	.338*		
Q9_FairWay	.314*	.342*	.617**	

Notes

***Corrolation significant at the 0.01 level (2 tailed)*

**Corrolation significant at the 0.05 level (2-tailed)*

Digital Testing Construct from the 2017 pilot

Table G

Reliability of Digital Acceptance Construct Items in 2016 pilot before deleting Q5

Measure	N	Corrected Item-Total Correlation	Cronbach's Alpha if Item deleted
Q1_GoodWay	365	.698	.737
Q2_FairWay	373	.643	.764
Q5_ReCRatioPrefer	330	.683	.746
Q6_ReCTradBetterScore	355	.528	.814
DigiAcceptance2017	280		.815

Table H

2017 Pattern Matrix using merged data of EE and paper-based (thus approx. 350 responses)

	Digital Testing Acceptance	Factor 2
Q1_GoodWay	.**826**	.062
Q2_FairWay	**.702**	.007
Q3_MCQEasier	-.033	-.433
Q4_ReCRatio Right*	-.535	.135
Q5_ReCRatioPrefer*	.**839**	.072
Q6_ReCTradBetterScore*	.**585**	-.032
Q7_TutGoodPrep	-.071	**.568**

Notes

Factor loadings above .400 were bolded.

*Q4, Q5 and Q6 was recoded first, so that a high score among all items would mean a high acceptance towards digital testing.

Table I

2017 Pattern Matrix chosen questions for digital acceptance

	Digital Testing Acceptance
Q1_GoodWay	.809
Q2_FairWay	.725
Q5_ReCRatioPrefer*	.782
Q6_ReCTradBetterScore*	.585

*Q5 and Q6 was recoded first and only before conducting this factor analysis, so that a high score among all items would mean a high acceptance towards digital testing.

Appendix 11: Open questions coding scheme

Firstly both pilots were scanned and checked for keywords.
Keyword/phrases that seemed to appeared often were: dislike; unfair; stress; "one error" ; "Method is more important" ; fraud; "Show more calculations"; "Mistakes"; "points for steps"; "not a good idea" ; "diagnostic testing", "not careful enough". There were too many categories, some categories were combined:

Before	After
Dislike digital testing on laptop	Digital testing is disliked and/or is stressful
Digital testing Maths is a bad idea	
Stress with Q left open	
Digital Testing causes a lot of stress	
Digital Testing is unfair	Digital is unfair; Cannot show what you know
Paper exam reflects knowledge better	
Paper is more fair than digital	
Every step is a point which you can gain, but not with final answer; cannot view steps	
One typing error in final answer results in losing all the marks	One error result in losing all marks
One calculation error results in losing all the marks	
Process/Method is more important than the answer	Method is more important than answer
Preference for being able to (digitally)show more steps	Preference: Show more steps digitally
More handwritten questions in the exam	Comments on design/ratio of the exam
MCQ is fine for basic questions	
(semi-)Positive remark	Positive or OK remark
Technical/practical suggestions	Practical suggestions for exam
Would like a person to check work	In addition: Human checking
Costs time	Chromebook for Maths is laborious
Difficult to type on a chromebook	
Fraud concerns	Fraud concerns

For interrater reliability, the flowing examples were given:

Code number	Coding Scheme Examples
1	**Digital testing is worse/disliked and/or is stressful** Words like: stress, bad idea, dislike, paper is better, anything that mentions the pop-up that warns not everything is filled in.
2	**Digital is unfair; Cannot show what you know**

Words like: unfair, [a[er exam reflects knowledge better, paper is more fair, every step is a point you can gain, capabilities

3	**One error result in losing all marks** words like: typing errors, or small mistakes or calculation errors.
4	**Method is more important than answer** Words like: calculation, process, method, working out is more important
5	**More steps digitally** words like : inbetween steps, checkpoints, more steps
6	**Comments on design/ratio of the exam** Comments like: that there should be more hand written questions, MCQ should be used for basic questions
7	**Positive or OK remark** workd like : OK, fine, I like it
8	**Practical suggestions for exam venue** suggestions such as: more plugs, more space, noise disturbance, extra desk
9	**Human checking also** words such as: person should also check
10	**Chromebook for Maths is laborious** words such as: input difficult, hard to type, no numpad, laptop not easy to use
11	**Time consuming** words like: waste time, takes a long time
12	**Fraud concerns** words like : cheating or fraud
13	**Good for diagnostic exams** words like : diagnostic
98	***Other***
99	***None/Don't Know/NA***

Inter-reliability calculations

Table L

Example of Inter-rater coding.

Comment	1	2	3	4	5	6	7	8	9	10	11	12	13	98	99
There was one mistake…etc etc…			B			D				B				A	

Table M

Categories of Coding of 59 comments across in the 2016 and 2017 exams.

Categories	1	2	3	4	5	6	7	8	9	10	11	12	13	98	99
Both	12	19	17	9	5	6	3	7	3	10	3	5	2	6	1
AOnly	2	2	0	2	0	0	1	1	0	1	0	0	0	1	0
DOnly	1	5	0	5	1	4	0	0	0	3	2	0	0	2	0
ATotal	14	21	17	11	5	6	4	8	3	11	3	5	2	7	1
DTotal	13	24	17	14	6	10	3	6	3	13	5	5	2	8	1

Note. Both means agreement among raters. Raters are identified with letters A and D. A identifies the researcher.

Table M

Interrater Table for both pilots

2016 and 2017 pilot responses		Rater D		Total
		Rated	Did not Rate	
Rater A	Rated	111	10	121
	Did not rate	23	741	764
Total		134	751	885

Cohen's Kappa calculations

$K = \frac{Po-Pe}{1-Pe}$

Po = $\frac{111+745}{885}$ = 0.967 #Agreement between raters

Pe = PRated +PNotRated

PRated = $\frac{134}{885}.\frac{121}{885} = 0.0207$ #Chance of rating same

PNotRated = $\frac{764}{885}.\frac{751}{885} = 0.7326$ #Chance of both not rating

Pe = 0.0194 + 0.740 = 0.7532 #Total chance

K = $\frac{0.967-0.7532}{1-0.7532} = 0.87$ #Cohen's Kappa

Appendix 12: Evaluation questions analysis in detail

Table J

Detailed Evaluation Reponses of all students in the 2016 exam.

Measure	1	2	3	4	5	Mean	SD
Q2_ExpMLP	0	3	18	**26**	8	3.71	.786
Q3_InputMLP	2	4	7	**21**	20	3.98	1.073
Q4_DigiOnlyFair	**31**	18	4	2	0	1.58	.786
Q5_DigiOnlyAdv	**24**	14	15	2	1	1.95	1.018
Q6_DigiOnlyNeat	13	**19**	13	8	2	2.40	1.116
Q7_CheckAns	6	10	17	**19**	3	3.05	1.096
Q8_GoodWay	12	**19**	16	6	2	2.40	1.065
Q9_FairWay	10	16	**18**	11	0	2.55	1.015
Q10_OwnDevice	2	2	4	22	**25**	4.20	.989
Q11_FraudProof	5	14	14	**15**	7	3.09	1.191

Note. Mode in boldface. Q2 to Q11 consisted of scale 1 = totally disagree to 5 = totally agree.

Table K

Detailed Evaluation Reponses of all students in the 2017 exam.

Measure	N	1	2	3	4	5	Mean	SD
Q1_GoodWay	365	43	101	74	**122**	25	2.96	1.17
Q2_FairWay	373	48	**113**	95	102	15	2.79	1.10
Q3_MCQEasier	362	24	119	**127**	76	16	2.84	.98
Q4_RatioRight_Total*	366	25	**130**	82	102	27	2.93	1.10
Q5_RatioPrefer_Total**	330	5	66	**110**	100	49	3.37	1.01
Q6_TradBetterScore	355	15	77	**160**	79	24	3.06	.93
Q7_TutGoodPrep	351	**264**	87	/	/	/	1.25	.43
*Q4_RatioRight_EE	44	**-**	**31**	-	13	-	2.6	0.92
*Q4_RatioRight_Other	322	25	**99**	82	89	27	3	1.11
**Q5_RatioPrefer_EE	30	0	**0**	14	12	4		
**Q5_RatioPrefer_Other	300	5	**66**	96	88	45		

Note. Mode in boldface. Q1-Q4 and Q6 were on the scale: 1 = totally disagree to 5 = totally agree. Q5 was scale (MCQ – W) 1 = 100 - 0, 2 = 75 - 25, 3 = 50 - 50, 4 = 25 - 75, 5 = 0 – 100.Q7 was 1= yes, 2 = no.

Table L

Detailed evaluation Reponses in the 2017 exam for Engineering students.

Measure	N	1	2	3	4	5	Mean	SD
Q8_InputMLP	44	1	8	4	15	**16**	3.84	1.180
Q9_WIFIGood	44	0	0	1	10	**33**	4.73	.499
Q10_NoTechIssues	44	1	1	2	9	**31**	4.55	.875
Q12_CheckAns	43	**12**	11	5	9	6	2.67	1.443
Q13_ShowScore	44	11	**13**	9	8	3	2.52	1.248
Q14_FraudProof	44	5	6	**20**	11	2	2.98	1.023
Q15_PreferDigitalExam	44	8	**18**	14	4	0	2.32	.883

Note. Questions were on the scale: 1 = totally disagree to 5 = totally agree.