UNIVERSITY OF GOTHENBURG

DEPARTMENT OF EDUCATION, COMMUNICATION & LEARNING

ON THE COMPARABILITY OF PAPER-BASED AND COMPUTER-BASED ENGLISH READING COMPREHENSION TESTS

A Study of High-Stakes English Reading Assessment

Lena Asp

Thesis:	30 higher education credits
Program and/or course:	International Master’s Programme in IT & Learning
Level:	Second Cycle
Semester/year:	Autumn term 2018
Supervisor:	Monica Rosén
Examiner:	Markus Nivala
Report no:	HT18-2920-001-PDA699

Abstract

Thesis:	30 higher education credits
Program and/or course:	International Master's Programme in IT & Learning
Level:	Second Cycle
Semester/year:	Autumn term 2018
Supervisor:	Monica Rosén
Examiner:	Markus Nivala
Report No:	HT18-2920-001-PDA699
Keywords:	Computer-based test; paper-based test; score comparability; language assessment; high-stakes testing; delivery mode

Purpose: The Swedish National Test will be digitalised by 2022 according to the government bill "Prop.2017/18:14. This study addressed the transition of high-stakes English reading comprehension tests from paper-based to computer-based test and examined whether a new delivery mode from an assessment and a measurement point of view can be regarded as equivalent to the traditional paper-based test.

Theory: The study was based on Classical test theory and Language testing theory.

Method: The empirical basis for this quasi-experimental study consisted of data from large-scale pilot studies carried out for future English high-stakes tests for year 9 in Swedish compulsory school. The database included possibilities to address the question of comparability in a quasi-experimental study where data from schools which were administered the test on computers can be compared to schools which were administered the test on paper. A total of 1275 English students in year 9 participated in the study. As the groups participating were not necessarily equal and comparable before this study, variables to control for some of the initial differences were added. An independent T-test indicated that the groups were comparable in terms of grade in English.

Results: On test items requiring constructed responses students in the computer-based test group gave on average lengthier responses compared to the paper-based group, but the difference did not result in better performance with higher scores. Regression analyses revealed that the delivery mode had no general effect on the scores. Results indicated however that boys' scores increased when the test was computer-based, more so on the shorter reading comprehension tasks than the extensive reading comprehension task. Findings in this study are discussed in connection to future research.

Foreword

Around the year 2008 schools in Sweden began use computers more and more in education. The school I worked at lent each student his or her own device and I found my job as a language teacher changing. The use of the internet brought my language classroom out into the world where the languages were actually spoken and new ways of teaching were introduced. I got an increased interest into the possibilities and the implications for new ways of language education which were made possible by ICT and I began the *International Master's programme on IT and Learning* at the University of Gothenburg. This lead to my current position as a test developer for the NAFS-project and more specifically, the work I do regarding the transition of delivery mode from a paper-based test into a computer-based test. My master's thesis has been inspired by the daily work we carry out within the NAFS-project and more precisely, the transition from paper-based to computer-based tests I am involved in. I would like to thank Dorte Velling Pedersen for valuable advice and support. Additionally, I would like to thank Stefan Johansson for wise comments on statistical analysis. A warm and very special thanks to my supervisor Monica Rosén who patiently has inspired me and given me insightful supervision, and introduced me to world of statistics. And last but not least, I am so grateful for the unconditional support from my family.

Lena Asp

September, 20018

List of Abbreviations

CAT	Computer-adaptive test
CALL	Computer-assisted language learning
CB	Computer-based
CLIL	Content-based language integrated learning
CR	Constructed response items
EFL	English as Foreign Language
ELL	English Language Learners
NAE	National Agency for Education
NAFS	National Assessment of Foreign Languages
PB	Paper-based
SA	Short answer item
SLA	Second language acquisition
SR	Selected response items

List of Tables

List of Figures

Table of content

1 Introduction

This thesis is concerned with the digitalization of the national tests in English, and whether this new delivery mode from an assessment and a measurement point of view can be regarded as equivalent to the traditional paper-based test.

Digitalisation is spreading through society and no area is left unaffected. Computer technology was used in language testing already in the 1960s for its efficiency (Chapelle & Voss, 2016) and to facilitate delivery and administration of tests (Chalhoub-Deville, 2001; Davey, 2011). As schools are a reflection of society, computers have been used in education for more than a decade and will continue to play an important role in language learning (Chapelle, 2007). In several Swedish schools, each student has access to a computer or a tablet to use in school and education has changed accordingly, making 21st century skills (to be explained in section 3.7) more disseminated. In light of the digitalisation of society, the Swedish government has decided that the national tests in schools will move into a digital environment. Thus, large-scale high-stakes tests are changing delivery mode from paper-based tests (PB) to computer-based tests (CB). When such a change occurs and new formats will be used, the new test can fail to assess the same ability and construct as before (Bachman, 2000; Alderson, 2000; Chapelle & Douglas, 2006, Chapelle, 2010; Douglas, 2000; Douglas & Hegelheimer, 2007).

Subgroups might perform differently on tests depending on delivery mode, and, according to Douglas & Hegelheimer (2007), very little research has been done in this area. Douglas (2010) believes that there is no need to question if modern technology should be used for language testing. According to him, it is already happening and will remain the delivery mode in the future. However, Douglas emphasises the need to investigate if performance differs with different delivery modes. He also asks if the development of new tasks is affected by technology. Likewise, Douglas asks if the definition of the language ability construct being measured, is affected by technology. Hence, he explains that language test developers should investigate if test takers' performance of the test will differ when new technology is introduced (Douglas, 2010). Thus, it is of great importance to empirically investigate whether the change of delivery format produces different results among the students.

1.1 The context of the study

The Swedish government announced in the government bill "Prop.2017/18:14" that national tests should be digitalized by 2022. Consequently, Skolverket, the Swedish National Agency for Education (NAE) was commissioned to execute this bill for all subjects tested. Hence, there is an ongoing project responsible for the transition from paper-based to computer-based testing of the large-scale, high-stakes national tests in Sweden. The transition is following a migratory strategy (Ripley, 2009) where the items used in paper-based tests are transferred into a screen-based environment. On commission of the Swedish National Agency for Education, the NAFS-project (National Assessment of Foreign Languages) at the Department of Education and Special Education at the University of Gothenburg, develops mandatory national tests of English, national assessment material for French, German and Spanish. The NAFS-project is conducting research which aims to investigate the transition of paper-based test into computer-based test to ensure equal validity and reliability.

This paper is part of the abovementioned process and the on-going work within the NAFS-project. The empirical basis consists of data from large-scale pilot studies carried out for future English high-stakes tests for year 9 in compulsory school. The database includes possibilities to address the question of comparability in a quasi-experimental study where data from schools which were administered the test on computers can be compared to schools which were administered the test on paper. There has been no random assignment to treatment condition as required in an experimental design, but the database includes some variables that to certain degree can compensate for this weakness. These field tests cannot be regarded as exactly equivalent to the high-stakes tests as they do not have any impact on students' grades, but the test tasks are the same. The main focus of this study is to investigate the comparability of the scores from the paper-based test and the computer-based test.

1.2 Overall aim

With the transition from paper-based test to computer-based test in mind is it important to gain knowledge into possible differences which may occur between the two delivery modes. The assumption that tests in different delivery modes are equal cannot be made without comparisons of students' achievements on these two delivery modes. It is vital to gain knowledge of how tasks and items are received by the test takers and whether the scores are comparable. It is also important to learn if differences in layout as well as reading on paper and on screen affect the test takers. It is also important to investigate group differences in this new delivery mode, so that they remain comparable to those known from the paper-based format. If different groups of test takers benefit from the new mode then there is reason to believe that the delivery format cause bias. Paper-based tests can be seen as the golden standard as it is and has been used for a considerable period of time and this is the standpoint taken in this thesis. However, as test takers become more familiar with computers and gain an increased digital literacy, tests delivered on paper might cause bias for other groups as well as other kinds of problems might arise. Last but not least, one need to ask if a test in this new delivery mode assesses the same construct as was tested with the previous testing methods. These initial research questions will be presented in more detail after the literature review.

1.3 Structure

First, this thesis begins with a presentation of the Swedish national tests and in more detail, the tests of English for year 9. This is followed by a theoretical background relevant to the field of high-stakes language assessment. Next follows a presentation of concepts relevant to the field of high-stakes assessment, and of research focusing the reliability and validity of results of achievement test using different delivery modes. The fourth chapter contains a literature review that focuses on more recent research on score comparability between paper-based and computer-based tests and the chapter ends with a presentation of the more precise research questions of this thesis. Thereafter follows a methodological chapter presenting the design, the data and the analytical methods used in this thesis. The sixth chapter presents the analyses and, finally, the last chapter discusses the results and conclusions in the light of previous research.

2 National tests in Sweden

This chapter gives a short background to high-stakes testing in Sweden and in particular, the testing of English for year 9.

2.1 Assessment in a historical context

Since the 1980s, the so called standardised central tests were given to assess students' skills and abilities in relation to a norm related grading system. The tests focused on receptive skills as well as on productive and interactive skills and intercultural communicative competence (Erickson, 2017). In 1994, this grading system changed into a goal- and criterion-reference system. The purpose of the tests was twofold; firstly to interpret and clarify the curriculum. Secondly, it should ascertain equity within the Swedish school system and aid teachers in the summative assessment (Erickson, 2017). However, not the whole syllabus is covered by the national assessment materials as they are merely advisory and not final examinations (Erickson, 2017). The latest curricula for both secondary school and upper-secondary school came in 2011. There is today, in 2018, instructions that teachers when grading should pay special attention to the test results and combine the observations they have carried out in class with the test result for each student. This new instruction signals that the results on the national tests should receive more attention from the grading teacher, and thus, the tests have become even more important. The main purpose with the national tests is currently to support equal and fair grading of the student's proficiency of the subject[1].

Grading is done on basis of the national curriculum and the national syllabus for each subject and the core content, objectives, and performance standards form the knowledge requirements. Teachers should as mentioned above also pay special attention to the result form the national test. Grades are awarded from the 6th year based on the individual student's achievement of the knowledge requirement. The Swedish grading system uses a 6-graded scale where grades are assigned as letters, A through F. A is Exemplary, C is Good, E is Acceptable, and F is Fail, not passed.

The curricula for English and Modern Languages (Spanish, German and French) have been developed with the Common European Framework of Reference (CEFR) as an important reference (Skolverket, 2017a, Erickson 2017). Since the new curricula in 2011 mentioned above, the curricula for languages have been further harmonized towards the Common European Framework of References. The curriculum and syllabus for English thus form the basis for the construction of national English language assessments.

This paper is limited to investigate the reading comprehension part of the assessment of English as a foreign language for year 9 in compulsory school. The school subjects of English and Modern Languages have however the same general language proficiency requirements. The assessment for year 9 aims to correspond to the CEFR-level B1.1. The CEFR-scale has an illustrative scale with six levels, from A1 to C2. The A level is the "English Basic User", B "English Independent User", and C "Proficient English User". As the CEFR framework is based on an action-oriented approach it looks upon the learners as social agents who have tasks to "accomplish in a given set of circumstances" (Council of Europe, 2018, p.9). Thus, the proficiency levels use "Can DO" descriptors to define each level. The descriptors focus on qualitative aspects of language use such as range, accuracy, fluency, interaction, and coherence (Skolverket, 2017a). The table 2.1 describes the relationship between the proficiency levels in CEFR and in what school year each level is expected to be reached by the students.

[1] https://www.skolverket.se/a-o/landningssidor-a-o/nationella-prov

Table 2.1 CEFR-levels for English in the Swedish educational system

CEFR level	**A 1.1**	**A 1.2**	**A 2.1**	**A 2.2**	**B 1.1**	**B 1.2**	**B 2.1**	**B 2.2**
STEP		1	2	3	4	5	6	7
compulsory school			year 6		year 9			
Upper-secondary school						En5	En6	En7

All the national tests follow a specific framework developed by the National Agency for Education in collaboration with the universities that develop the different tests (Skolverket, 2017b). This framework was commissioned by the Swedish government in the bill “Prop 2017/18:14” to secure high quality and the highest possible credibility of the national tests.

2.2 National assessment of English

As mentioned in the previous section, the aims of high-stakes tests are to support equal and fair assessment. The national tests of English as a foreign language (EFL) is administered every year nationwide in Sweden by the Swedish National Agency for Education, to approximately 100,000 test-takers, at the same day and time. The Swedish national tests are used to make high-stakes decisions such as awarding grades to students and it could also be used as a measure of school performance and national achievement. The language tests are proficiency tests aimed to assess what the pupil knows and can do in the world outside of school. This is different from the regular achievement tests given by teachers in class which aim to assess the learning outcome of a certain objective taught in class, e.g. a month’s work or a chapter in a course book (Council of Europe, 2018).

It is a mandatory requirement for students to take the English national tests at secondary education (in year 6 and 9, student aged 12 and 15), and one test upon completing the final course of English at upper secondary education (students aged 16 and 17). As mentioned before, the project *Nationella Prov i Främmande Språk* (NAFS, National Assessment of Foreign Languages) at the Department of Education and Special Education at the University of Gothenburg, is commissioned by the Swedish National Agency for Education to develop these mentioned high-stakes tests of English, and language assessment material of French, Spanish, and German.

2.3 The National English test

The complete national assessment material of English, tests both oral and written production and interaction, as well as receptive competences of reading and listening. The English test consists of three parts: A (focus Speaking), B (focus Reception – reading & listening), C (focus Writing).

The first part, A, is an oral test, assessing the test taker’s ability to express him/herself and interact in English with a partner. Each pair is given 15 to 20 minutes to complete the task. There are instructions and material for the task. In general, there is a warm up, followed by a discussion where cards or sheets with topics to talk about are presented to the test takers. Finally, the test ends with a more argumentative discussion with statements or questions the test takers should discuss in more detail. The conversation is often recorded by the teacher to enable assessment of other teachers as well. This task is assessed by a teacher who is aided by extensive guidelines and authentic samples of benchmarks.

The second part, B, assesses the receptive ability of reading and listening. The two sub-parts consist of different texts and listening items. The test takers start with the reading comprehension and get 90 minutes to complete this part. After this, the test takers have a break for approximately 30 minutes before they start with the second part, the listening comprehension. This part is approximately 50 minutes.

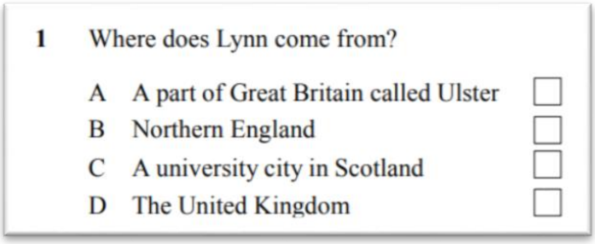
It is important to secure that no test taker has advantages due to prior knowledge. Important principles for the construction of test items and the scoring guides are therefore that the responses asked for should be found in the texts or in the audio file and no test taker should be rewarded for giving further facts or information.

Part B is composed with different tasks and items which use various response formats to ascertain the reliability of the score and the validity of the construct. Chapelle & Douglas (2006) discuss the importance of using different kinds of formats; multiple choice items and other fixed formats combined with different kinds of open ended responses. Tests of language abilities should use different response formats in order to balance possible influences of any single response formats. This is also supported by Alderson (2000) who argues that different aspects of the construct could and need to be assessed using different methods:

> *It is now generally accepted that it is inadequate to measure the understanding of text by only one method, and that objective methods can usefully be supplemented by more subjectively evaluated techniques. Good reading tests are likely to employ a number of different techniques, possibly even on the same text, but certainly across the range of texts tested* (p. 206).

Erickson (2009b), the former manager and principal researcher in the NAFS-project, also emphasizes the importance of a variation of texts and formats of responses to ascertain the validity in assessments. The different formats used in the English national test include selected response (SR) items, matching items, short answer (SA) items, and constructed response (CR) items where a more extensive response of one to several sentences should be entered. Figures 2.1 – 2.6 show examples of formats used in the English tests. The *selected response items* include three response formats; multiple choice, multiple matching and multiple cloze.

In Fig. 2.1 a multiple choice format is shown. The test taker reads a text or listens to an audio file and is given questions on the content and has to choose one correct alternative of three to four given alternatives.

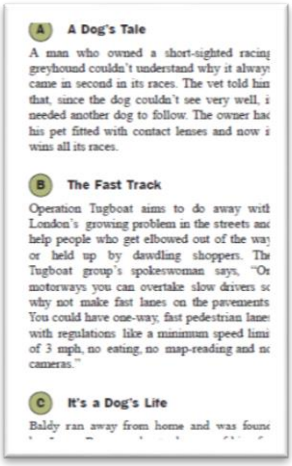


1 Where does Lynn come from?

A A part of Great Britain called Ulster □
B Northern England □
C A university city in Scotland □
D The United Kingdom □

Fig. 2.1 Multiple choice item.

The multiple matching format is displayed in Fig. 2.2. The test taker reads a text, often several shorter texts which are numbered. Then the test taker comments the statements by choosing one of the texts that fits the statement. The format is also used for assessing listening comprehension.



A **A Dog's Tale**

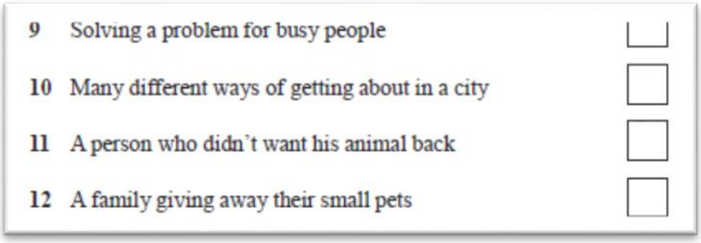
A man who owned a short-sighted racing greyhound couldn't understand why it always came in second in its races. The vet told him that, since the dog couldn't see very well, it needed another dog to follow. The owner had his pet fitted with contact lenses and now it wins all its races.

B **The Fast Track**

Operation Tugboat aims to do away with London's growing problem in the streets and help people who get elbowed out of the way or held up by dawdling shoppers. The Tugboat group's spokeswoman says, "On motorways you can overtake slow drivers so why not make fast lanes on the pavements. You could have one-way, fast pedestrian lanes with regulations like a minimum speed limit of 3 mph, no eating, no map-reading and no cameras."

C **It's a Dog's Life**

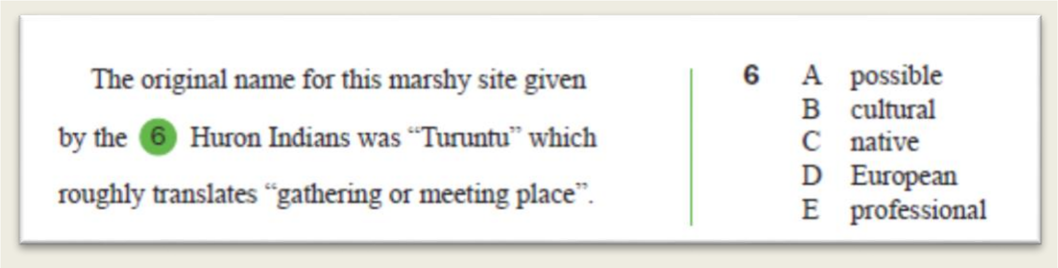
Baldy ran away from home and was found



9 Solving a problem for busy people □
10 Many different ways of getting about in a city □
11 A person who didn't want his animal back □
12 A family giving away their small pets □

Fig. 2.2 Multiple matching item.

Fig. 2.3 illustrates a so called multiple cloze format. The test taker reads the text in which there are some gaps with missing words and has to decide which of the given alternatives fits the context.



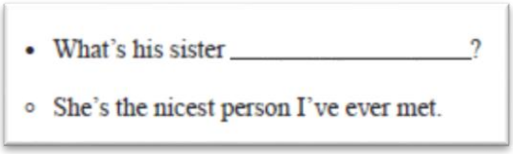
The original name for this marshy site given
by the 6 Huron Indians was "Turuntu" which
roughly translates "gathering or meeting place".

6 A possible
B cultural
C native
D European
E professional

Fig. 2.3 Multiple cloze item.

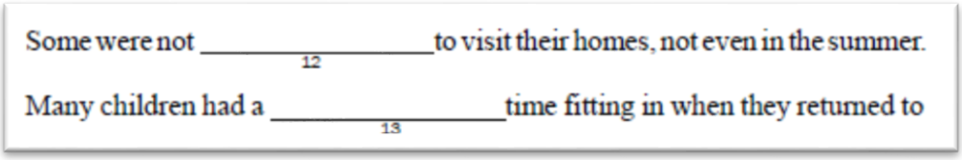
Among the *constructed response items*, the productive gap and open ended formats are used.

The productive gap test format is displayed in Fig. 2.4 and 2.5. This task consists of either shorter sentences, dialogues or longer texts. The test taker reads the text and fills in only one word in each gap.



- What's his sister ________________?
- She's the nicest person I've ever met.

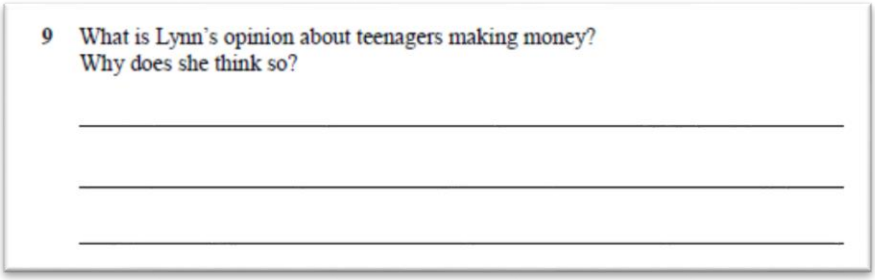
Fig. 2.4 Gap item – dialogue.



Some were not ________________ to visit their homes, not even in the summer.
12
Many children had a ________________ time fitting in when they returned to
13

Fig. 2.5 Gap item – text.

An open ended, constructed response format is displayed in Fig. 2.6. The test taker reads a text or listens to an audio file and writes a response to the questions on that text.



9 What is Lynn's opinion about teenagers making money?
Why does she think so?

____

____

____

Fig. 2.6 Open ended, constructed response.

Multiple choice items are, according to Boyd & Taylor (2016), reliable but a test cannot merely consist of multiple choice items, it should also consist of items that are meaningful in a context outside the test. Hence, there is a need for a balance between multiple choice items and open ended, constructed responses where the test taker has to produce a response. However, there has to be a mix of items, formats and of different authenticity levels, and the more authentic the items are the more the reliability could be affected negatively. This is explained by Boyd and Taylor:

> *There is a tension between authenticity and reliability. The closer an assessment task replicates a real life exchange or communicative event, the more this very strong content validity introduces variables that inevitably impede efforts to ensure robust reliability* (p. 40).

In the English national tests, some items are dichotomous meaning that responses are scored either 0 for incorrect or 1 for correct responses. Other items include grading of responses, that is for example when items are being scored 0, 1 or 2 (or more). In the constructive response items, the test takers write a response which requires a subjective assessment from the teacher. This assessment is carried out on basis of benchmarks supplied in the teacher guidelines to aid the teacher in the scoring process. The benchmarks are exemplified with authentic test taker comments from large-scale pilot tests. The scores from the two parts of the English reading and listening comprehension tests are combined into a single score for this receptive part, thus resulting in a sub-score for reading comprehension and a sub-score for listening comprehension

Finally, part C assesses the test taker's writing proficiency. The test taker is given a topic which includes some inspirational information and prompts. The time frame is 80 minutes to complete the task in year 9. For the assessment and scoring, the teacher is provided in the teacher guidelines with exemplifying benchmarks of authentic responses collected from large-scale pilot tests of the task.

It is the teachers at the local schools who assess and score their student responses. The results from all three parts are combined into a final test score which the grading teacher should take in consideration (in addition to her/his other observations of students' performance) for the grading process at the end of compulsory school. Then, all results in the test are reported to the National Agency for Education by the schools. Furthermore, a selection of tests on basis of the test takers birth date are sent back to the NAFS-project for further analysis and quality assurance.

The next section will briefly describe the process of test development of the English national tests.

2.4 Task development process within the NAFS-project

The development of tasks and items for the national language tests follow a standardized procedure. The NAFS-project employs test developers who all have a background as teachers of the specific language tested, there are also native speakers, language researchers, and language teacher educators. Within the NAFS-project test developers produce material for future tests. Different reference groups, including working teachers at different school levels, are also involved at different stages of the process.

New tasks, items and testlets are piloted at an early stage. This is conducted by smaller groups and then analyzed to investigate how the task works, if there are ambiguities or other problems. The tasks are then revised and tested in large-scales pilot tests by around 400 students for each item at different schools in Sweden in a random selection from a school register at the National Agency for Education. All material used are controlled and supervised by native speakers and language scholars. In the large-scale pilot tests, anchor items are included to ascertain the language ability of the test taker. Anchor items are and have been used to make the tasks and items comparable over time and across groups. All tasks are commented regularly in surveys by both teachers (TF- teacher feedback) and test takers (TTF – test taker feedback). The tests are sent back to the NAFS-project to be coded and assessed by trained professional raters. The results are imported into statistical software for further analysis of the data. Different qualitative and quantitative methods are used to analyze the result. The feedback from teachers and test takers are as vital in the process as the data from the results. The analysis will give information on reliability and other descriptive statistics and also whether certain items function better than others. The items which do not show reliable results are removed or adjusted and then pre-tested again. Only items which show high reliability and validity are included in the tests. All data from the large-scale pilot tests are stored in databases for further research.

All test items included in this study come from the NAFS database and are part of the pre-large-scales tests process described above. Due to confidentiality of the items no disclosure can be made. However, in the method section similar items will be exposed to exemplify the actual items in the study.

3 Theoretical background

The following chapter will give a conceptual framework with a theoretical background of language assessment. Firstly, the empirical foundations of the study are presented. This theoretical background includes concepts which are relevant in the field of assessment and more specifically identify key concepts relevant when tests change delivery mode. Secondly, literacy and digital literacy will be described. Finally, computer-based assessment will be presented.

3.1 Assessment

The outcome of the collection and processing of information about something meant to be inferred is called an assessment (Bachman and Palmer, 1996). According to Bachman and Palmer, an assessment is a designed systematic proceeding that can be replicated at later stages and it is grounded in a confirmable area of content, in this context the syllabus for English in Swedish schools. The quality of an assessment depends mainly on the reliability and validity of a test result. Bachman and Palmer explain that high validity follows when results are reliable and consistent with the targeted construct.

Large-scale assessments are tests administered to a large number of test takers, such as the students in the same year in a nation. The tests have different stakeholders, one group is the individuals or the students who will be affected by the decisions made on basis of the results, e.g. grades or admissions to schools and this could have major consequences for these stakeholders (Bachman and Palmer, 1996). Another group of stakeholders are the schools or organizations that are affected by the results of the test when it comes to funding and allocation of resources. The third group of stakeholders is the state or the country when its education system is evaluated via the tests.

Thus, high-stakes test are tests used to make decisions about the test takers, the schools or the nation whereas low-stakes tests have no significant consequences for either test takers, teachers or schools.

3.2 Construct

The construct is a theoretical concept denoting *what* a test is intended to measure (Douglas, 2010). Douglas points out that the construct needs to be clearly defined and it must be possible to prove that the test, the items included and the test result are relevant not merely to the construct but also in relation to the purpose of the test. Douglas is supported by Weigle (2002) who also argues that the definition of the construct is one of the corner stones in test development. The construct of a second language (L2) test is multidimensional and involves various interacting processes and components (Bachman & Palmer, 1996).

It has been argued that a change of delivery mode from paper-based-tests to computer-based-tests risks affecting the definition of the construct (Takala, Erickson, Figuera and Gustafsson, 2016). Thus, a new delivery mode brings on a need for research into construct validity (Sawaki, 2001; Davey, 2011; Kozma, 2009). An important research task is therefore to determine the extent to which computer-based tests and paper-based tests are equivalent and measure the same construct.

Chapelle and Douglas (2006) acknowledge that digital technology can provide other ways of testing that are not possible in a paper-based test. For example it is possible to measure the amount of time it takes for a test taker to give a response to an item. The quicker the response, the more probable it is that the language skills have been automatized (*Ibid* p.16). Pommerich (2004) explains how evaluations and analysis at item level will give understanding into the possible sources of mode differences. Such analyses can provide empirical knowledge of how test takers interact with items whether they are presented on paper or a computer interface.

Kyllonen (2009) goes a bit further in the discussion of digitalisation of tests in general, and sees the possibility to test other constructs than what has been possible with the paper based format. However, this would mean that new constructs may be measured. Still, for National tests in English, it is important that everything assessed is part of the national curricula and subject syllabus and therefore meant to be assessed.

3.3 Reliability and validity

Test scores have to be reliable. The concept of reliability refers to the overall consistency of the scores of an assessment, and the reproducibility of a score if the test would be taken a second time. However, since giving the test to the same students a second time is not feasible nor reasonable, other methods are needed to determine the reliability of the test. High reliability follows when the number of test items are sufficient, and when the inter-item correlations are high. Ideally, a good test includes enough test items to cover all ability levels in the intended construct.

The concept of validity refers to how well the tests represents its construct. Besides a clear definition of the intended construct and the content of the tasks, validity also includes how well the operationalized tasks manage to measure the construct. Construct-validity is the most paramount factor when it comes to securing a sound assessment according to Takala *et al.* (2016). Consequently, a test has to make it possible to assess the construct. There are two main threats to validity; construct-irrelevant variance and construct-under-representation (Messick, 1989). Construct-irrelevant variance is when variables not related to the construct are measured. Construct-under-representation is when too little of the intended construct is covered by the items. Messick explains validity as follows:

> *"validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (*1989, p. 13)*."*

A result of construct under-representation or construct-irrelevant variance could be that inference done from the result are unsubstantiated, or it could show bias if groups of test takers are statistically favoured or disadvantaged for all the wrong reasons. Furthermore, validity is also affected by the inference done on the test result and that it is used in a way which gives reasonable consequences, hence consequence-validity (Messick, 1989, Erickson, 2009a, Bachman & Palmer, 1996, Douglas, 2010, Weigle (2002). As Douglas (2010) explains that:

> *validity is about collecting evidence to demonstrate that the interpretations and decisions we make on the basis of test performance are justified - to do so we must focus on the ability(ies) the test is intended to measure and the decisions we wish to make on the basis of the test performance* (p.26).

Validity and reliability are related in such a way that high reliability is a necessary but insufficient prerequisite for validity. This means that high reliability does not automatically ensure high validity. Low reliability however, does always imply low validity. This study has its primary focus on issues related to reliability and less on other aspects of validity.

3.4 Validation

The concept of validation refers to the process of determining whether a tests assesses the intended construct. It is vital that high-stakes tests maintain high quality, measure its intended construct and that the result are inferred correctly as test results will have an influence on the test taker, the teacher and the school system (Douglas, 2010). Douglas explains that the process of colleting proof that the construct intended is assessed is known as validation:

> *Validation is a process of gathering evidence to support the claim that a test measures certain abilities or attributes in certain contexts for certain purposes.* (*Ibid* p. 257).

Therefore, to secure that tests actually assess what they are intended to, there is a need for continuous construct validation, that is analyses of to what extent the test reflects the construct. This is exemplified by Weigle (2002) when explaining that a writing test should first of all assess written production and no other ability. A way to validate the test is to look at the consequential validity of the test. If a subgroup of test takers underperform on the test this could indicate that a test is not measuring its construct. In short, correct interpretations of the test result have to ensure that the inference lead to fair decisions (*Ibid*). Another way to validate tests is to investigate their relation to other tests of the same construct, and where high correlations are the desired outcome.

In view of the above, it is apparent that validation is important for each test which is transitioned form paper to computer (Russell *et al*., 2011).This fact is supported by Weir (2005) who argues the need for studies to secure validity. Weigle (2002) emphasizes the need of investigation into whether the test takers use different cognitive processes depending on delivery mode, computer- or paper-based, and also to gather empirics about the test takers' computer habits and computer familiarity (*Ibid* p. 235).

Not only items have to be validated throughout the process but also the function of the platform used to deliver and administer a computer-based test (O'Sullivan, 2015). Pommerich (2004) claims that as technology keeps on evolving new studies into digital platforms have to be conducted where researchers make comparisons between computer-based tests and the equivalent paper-based tests.

> *Item level evaluations can also provide insights into sources of mode differences and can help develop an understanding of how examinees interact with item features when presented in a test booklet versus a computer interface* (*Ibid* p. 4).

Therefore, not only analyses into computer- and paper-based delivery modes are needed, it is also important to investigate the various interfaces of computer-based tests depending on digital devices, e.g. tablets or computers. According to Pommerich (2004), the delivery mode of a test, the computer interface, will change depending on the platform and items will be perceived differently by the test takers. Consequently, it might not be possible to assume the comparability of a test administered on a digital platform with the same items delivered with another digital interface. Hence, there is a constant need to evaluate items and digital solutions since the computer-based mode will develop and change over time with new technological evolution (O'Sullivan, 2015).

The transition from paper-based delivery mode into computer-based tests will follow a structured process to ascertain reliability and validity of the test. Lindqvist (2012) saw in her empirical study three steps when making tests computer-based. According to Lindqvist, the first step is when the paper test is delivered on a screen and the pen has been replaced with a keyboard. The second step is when new formats and types of items are used and there is a change in the complete procedure of a test, its format, and its assessment. Then, the final, third step is when ICT-literacy is included in the assessment, which means that a new construct will be assessed.

To conclude, validation is an ongoing process which has to continuously look at all the different factors mentioned above. There is a constant need for validation of items and digital solutions.

3.5 Equality and Fairness

The main purpose of the national test is to support an equal and fair assessment of the test takers' proficiency and consequently fair grading (Skolverket, 2017b). To ascertain this, it is important to know the construct of the test according to Gipps & Stobart (2009). As mentioned previously, it is important that the tests allow the test takers to show their language abilities which is done through a variation of tasks and items so no test taker has advantages or disadvantages. Assessment has to provide results that are generalizable beyond the assessment itself otherwise the inferences made on basis of the assessment are not fair (Bachman & Palmer, 1996).

Klapp (2015) investigated the relationship between the grades for the subjects Swedish, English and Mathematics at the end of compulsory school in Sweden, and the result on the national tests, to see if there was some other factor common for all three subjects involved in the grading process. She points out that girls have an advantage over boys with regard to grading. Her studies found that there is a medium effect size of .51 in favour of girls (p.27). Her results show a variance in the grades that could not be explained by the results on the national tests. Thus, Klapp hypothesized that other factors which were not assessed in the national tests were involved in grading. She found gender difference in favour of the girls (0.15). When looking at the grades for Swedish, the value was 0.25, for English 0.08, and for Mathematics 4.4% of the grade was explained by something else than the grade on the national test (*Ibid* p. 67). Klapp explains however, that there should be a difference between grades and results on national tests as these tests do not cover the whole curriculum of a subject.

According to Lindqvist (2012), equality and fairness can increase when the tests become computerized. Lindqvist explains that when information material, time frames etc. become homogenous within a digital platform for the computer-based tests equality will increase. With this in mind, standardization of the test environment will strengthen the equality aspect.

However, there is also a risk of weakened equality when tests become computerized. This could happen if the test result gets dependent on the level of digital literacy of the test takers, and/or differences in familiarity with taking tests on computers. This is discussed by Weigle (2002) who claims that there exists a technology gap or a digital divide:

> *While certain types of computer-based assessment involve only rudimentary computer skills such as clicking and dragging with a mouse, writing on the computer is an entirely different matter, involving complex keyboarding skills. […] Unless such skills are part of the construct (as they might be, for example, in a test for office workers) it is clearly inequitable to require students with weak or non-existent key-boarding skills to use a computer rather than a pen and paper on writing tests* (p. 237).

However, there are some signs indicating that the digital divide is decreasing, Chapelle & Voss (2016) imply that more test takers are now using digital devices in school and in their spare time. In 2017, 95 % of all Swedes over the age of 12 had access to an internet connection at home and nearly all students in schools, more than 98 % use internet daily (Internetstiftelsen, 2017). This would signify that the digital divide is getting smaller and possible bias when paper-based tests become computer-based might diminish. Nonetheless, there would be an unfamiliarity with performing high-stakes test on screen. This is discussed by Chapelle & Douglas (2006) who urge a way of counteracting the digital divide by using demo tests and tutorials for test takers to get used to computer-based tests. This is supported by Douglas (2010) when he explains how tutorial and demo test in advance of a test will diminish test takers inexperience and their nervousness for doing the test. Messick (1989) clarifies that if test takers have to put effort into navigating in the digital platform and its interface rather than answering the items this will lead to construct-irrelevant variance and thus decrease the validity of a test result.

In sum, tests need to be fair in order for the results of the tests to be reliable and valid. Fairness and equality also have to do with things prior to doing the test, it could be the digital infrastructure of the school, such as its digital equipment and the capacity of the internet connection. But there is also the test takers digital literacy, experience and familiarity with computer-based tests as abovementioned. All this can lead to bias and effect the reliability.

3.6 Authenticity, practicality and method effects

The concept of authenticity relates to the fact that tasks included in a test should be representative of the language usage test takers will need in a world outside the test (Weigle, 2002). In order to strengthen the validity and interpretations of the test, authenticity is something which should be strived for. This is pointed out by Douglas (2010):

> *It is difficult to produce authenticity but we must nevertheless make an effort to provide a context for language use in our test to help ensure that the interpretations we make of the test takers' participation will be valid* (p. 26).

Authenticity refers also to the everyday situations students encounter at school. As students nowadays, in schools and at home, write and edit their texts on computers or other digital devices, the authenticity of a computer-based test would increase. In the digital writing process students use specific tools built in for this purpose (e.g. spelling programs) and their familiarity with using computers has increased (Internetstiftelsen, 2017). So further possibilities offered in computer-based tests like spelling programs, and the possibility to get text read aloud, will affect the construct and might not be what the syllabus describes and what tests are intended to measure (Chapelle & Douglas, 2006). However, digital technology could also be used to mirror real-life tasks as a means to increase authenticity, e.g. if computer-based tests use contextual linked information via audio, film clips and web pages.

The concept of practicality of a test refers to the usefulness of a test in relation to the resources needed (Bachman & Palmer, 1996). When tests change mode and become computer-based this will have an impact on education, a washback effect. According to Hughes (2003), the change of mode from paper to computer could show construct-irrelevant difficulty or easiness depending on the test taker’s familiarity with computers and his or her digital literacy. This will have a washback effect on education.

Method effects entail that a test taker’s results may be affected by the mode used in the test. Chapelle & Douglas (2006) claim that different formats can fail to assess the same ability and they urge the need for studies into the effects of a change of mode, and also studies comparing method effects on different formats to ensure that the same test taken in different modes or whether different formats measure the same construct. Thus, when test takers take a computer-based test on a digital device the test might be displayed differently depending on the device used. In short, depending on whether the computer-based test uses responsive or static representation this could also have an impact on the equality of the test. Consequently, Pommerich (2004) claims that the harder it is to view an item on a screen compared to on paper the greater the risk of method effects. Pommerich notes that reading comprehensions in computer-based tests where the test takers have to scroll to read the whole text have greater method effects compared to a test in Mathematics. She also implies how there are substantial mode effects on tests when test takers have to write responses. Mode effects can arise if the test takers endurance or motivation is affected by the change in delivery mode.

3.7 Literacy and reading on screen

Being able to read and write used to be a synonym for being literate. According to the Merriam-Webster online dictionary[3] there are two definitions of the word *literate*; the first definition is to be educated, cultured. The second is to be able to read and write. The word *literacy* is defined as the quality or state of being literate according to the same web-page. As previously stated, to be literate used to be that someone was able to read and to write. However, today this is no longer solely accurate. A society with an evolving and rapidly changing information and communication technology claims a new definition of literacy. Thus, the meaning of literacy is deictic since the meaning evolves over time and place. What we know to be literate today will not be valid in a few years and it will continue evolving (Leu, Kinzer, Coiro, Castek, Henry, 2017). According to Gee (2003), we need to be literate in several semiotic domains such as images, gestures, sound, graphs, equations, objects, symbols etc. Gee suggests that we also have to be able to learn new things in other domains through-out our lives. In this context there are

[3] https://www.merriam-webster.com/dictionary/literate

numerous concepts of literacy used today, just to mention a few: digital literacy (Knobel & Lankshear, 2008, 2015), and multiliteracies (Buckingham 2008, Gillen 2014).

In the PISA assessments 2000 and 2009 girls scored higher on reading comprehension compared to boys in all participating countries except from Israel and Peru (SOU 2012:10). A closer look at the Swedish results from PISA 2009 showed that girls scored higher on all reading skills. Furthermore, the highest difference was for the hardest competence (reflecting and interpreting) (*Ibid*: p. 94). Likewise, the group with the weakest readers consisted of 10% girls compared to 24% boys. Digital reading was tested for the first time in PISA 2009, and the difference of mean score between boys and girls was lower for the computer-based test than traditional reading on paper. The SOU (2016:25) argues that computer-based delivery of the Swedish national tests should decrease gender differences on the scores. An explanation drawn in this SOU is that boys are better at reading on a computer, digital reading, than reading on paper (p. 102).

Reading literacy does not only include the ability to read in any possible way, on paper or on screen, but also the ability to "handle, understand, and take advantage of the Internet environment" (Rasmusson, 2014b, p.14) She continues to state the difference between reading digital texts that use hyperlinks, audio etc. to the reading comprehension of traditional texts presented on a screen. This latter type of reading comprehension is linear like traditional reading and without hyperlinks. They are also presented in a layout similar to that on paper. More visual-spatial skills are supposed to be involved in digital reading compared to traditional reading on which males perform better than females (Rasmusson & Åberg-Bengtsson, 2015). They suggest that these visual spatial skills are further developed by computer game playing.

As said, our world is rapidly changing as new technologies are being developed and society in general as well as education need to adapt to this change. The New London Group with Cazden, Cope, Fairclough, Gee *et al*. (1996) presented the idea that there are various forms of texts made possible with information and multimedia technologies. The New London Group (1996) concluded that texts as multimodal products include multiple ways of communication, such as linguistic, visual, gestural, spatial, and audio (p. 83).

The European Council (2006) has produced a framework with eight key competences for lifelong learning which is believed to be a must for people in today's society. Number four in this list is digital competence. Thus, the ever evolving technology urges us to re-learn constantly. Accordingly, there is a need for a common framework of the meaning of being "digital savy in an increasingly globalised and digital world" (Vuorikari, Punie, Carretero, Van den Brande, 2016, p.3). As part of a lifelong learning programme in a digital world UNESCO (United Nations Educational, Scientific and Cultural Organization) has, in the book *Understanding Information Literacy: A Primer* (2008a), put down six points included in the"*21st Century Survival Literacies*" (*ibid* p.3*)*:

> *1. the Basic or Core functional literacy fluencies (competencies) of reading, writing, oralcy and numeracy*
> *2. Computer Literacy*
> *3. Media Literacy*
> *4. Distance Education and E-learning*
> *5. Cultural Literacy*
> *6. Information Literacy*

There is however, according to Leu *et al*. (2017), a "lack of valid, reliable, and practical assessment of new literacies to inform instruction and help students become better prepared for an online age of information and communication (*ibid:* p.11)". Leu *et al*. emphasize that there is an expanded definition of reading literacy which includes basic reading skills as well as higher-level digital reading skills. They argue that this re-definition of reading literacy will keep on changing as new technology evolves and continue to state that assessment needs to be dynamic and change together with the rapidly changing, deictic digital world. Or as stated the OECD (2018 p.10) *"reading must be considered across the varied ways in which citizens interact with text-based artefacts and how reading is part of life-long learning."*

Nevertheless, is reading on screen synonymous to reading on paper? To gain more information into this matter, Köpper *et al.* (2016) conducted a study and looked at studies from the 1980s which indicated higher eye strain when reading on screen. They looked into whether new technology with screens with better display resolution, luminance and contrast brought new findings. In their studies, they found no significant differences between proof-reading on screen or on paper. However, their participants showed significantly stronger symptoms of eyestrain after reading on screen. Rasmusson (2014a) came to the same conclusion, there are no clear results as to whether reading on screen and on paper are equivalent. Martin & Binkley (2009)., came to the conclusion that as high-stakes tests change delivery mode to being computer-based, the reading literacy gap which exists between girls and boys could decrease in favour for boys.

However, high-stakes tests cannot be allowed to change overnight, they have to evolve with the curricula which they are aimed to assess in a valid and reliable way.

3.8 Computer-based tests

Various types of computer-based tests are used. In linear computer-based tests, all test takers are given the same set of items whereas in computer adaptive test (CAT) an algorithm adapts the level of difficulty for each item the test taker is given and the number of questions depends on the responses given (Davey, 2011; Lindqvist, 2012). The early CAT-testing algorithms relied on a psychometric method called item response theory (IRT) (Chapelle & Voss, 2016). When tests are delivered on a computer new possibilities are available according to Chapelle and Voss (2016). They explain that other affordances are possible in computer-based testing because additional types of data can be gathered easily, such as time spent on each item. This can indicate if the language is automated and the language fluency of the test taker. However, computer-based tests may require additional time for completing a test compared to paper-based tests. Alderson (2000) explained the need for research into time allotment as tests change delivery mode. Bayazit & Aşkar (2012) reported from their study that the computer-based test of an equivalent paper-based test took longer time for the test takers to take.

Pommerich (2004) explains how reading tests, where there are texts viewed on a two-page spread on paper in a booklet but shown on a single computer screen, will cause issues. She argues that "The more complicated it is to present or take the test on computer, the greater the possibility of mode effects" (*Ibid.* pp. 3-4).

So, is the construct of English reading comprehension different when reading on paper compared to reading on screen? In 2009, PISA modified their framework to include electronic texts in their testing (OECD, 2018). Digital texts have other affordances and impose cognitive demands on their readers, e.g. reading on screen can be influenced by luminance contrast, and backlight, scrolling, and screen size. Rasmusson (2014a) found in her study that the performance on digital reading was better for the boys. She explained that the factor for this was that boys spent more time playing computer games compared to girls. This fact was empirically supported by Sylvén & Sundqvist (2012). They showed in their study that the more boys were involved in massively multiplayer gaming in their spare time, the better their English proficiency was. Consequently, the input of English that students get outside school, the so called extra-mural English (EE), has an impact on academic vocabulary at lower proficiency levels (Olsson & Sylvén, 2015). Olsson & Sylvén looked at students of English at upper secondary school and compared groups that got their education in English (EFL) in several subjects, CLIL (content and language integrated learning) and those who were taught in Swedish. Their findings show that "male CLIL students used a significantly larger proportion of academic vocabulary compared to all other groups" (*ibid* p.93).

The following chapter will focus on previous empirical research studies into comparability between paper-based and computer-based tests. The chapter will end with the aim and research questions for this thesis.

4 Literature Review

This literature review chapter will focus on some current empirical research studies on the topic of comparability between computer-based and paper-based tests. Building on previous research this chapter will end with the aim and more specified research questions for this thesis.

4.1 Search strategies

For the identification of literature, mainly the electronic search engines "Supersök", at the University of Gothenburg, as well as Education Research complete, ERIC, were used. In ERIC only two words were included: *language assessment* and *computer-based test.* A restriction to only include peer reviewed and fully accessible articles from the last 10 years was added. This resulted in 17 articles. To broaden the search and be able to find articles related to the key words an asterisk (*) was used. This enables searches for other versions and endings of the terms. This resulted in 21 articles. Out of these, only three articles were relevant for this thesis. In "Supersök" the terms *language assessment, computer-based test, paper-based test, test, validity,* and *delivery mode* were entered in the search field. Furthermore, only the past 10 years, the disciplines of pedagogy and language and literature were included in the search criteria. This resulted in 950 articles or books. The first quick review showed that most of the suggested studies on the list were irrelevant for the purpose of this study and the next step only the articles on language testing or comparability of paper-based test and computer-based test were selected. The abstracts of these were read and their reference lists studied in order to find articles of relevance for the current study. The final list contained 15 articles which are included in the literature review below, and listed in *Appendix 1* as well as in the list of references.

4.1.1 Inclusion and exclusion criteria

As mentioned, the criteria for choosing which articles to include in the review were that they had to be peer reviewed and published in a scientific journal within the last 10 years. However, not so much research into the comparability of L2, English reading comprehension tests, in a change of delivery mode were found. There were however articles of interest comparing the two test modes in the assessment of English writing and in other subjects which were included in the review of previous research.

4.2 Summary of articles

In a Norwegian experimental study, Mangen *et al.* (2013) looked at the L1 (first language) reading comprehension of texts in digital or print mode for 72 randomly selected students in grade 10 (aged 15 to 16). The students were split into two groups haphazardly. Group 1 read two texts in print and the other group the same texts but delivered on screen in a PDF-format. The length of the texts was between 1,400-1,600 words and they had illustrations and graphics. Prior to the tests the students took pre-tests in reading comprehension, word reading and vocabulary. A *t*-test revealed that there was no significant difference between the groups. Furthermore, there was no significant difference between narrative or expository texts regarding modality. However, for reading comprehension, statistically significant difference for better results for the paper-based test was reported (*p* =.025). The test scores were higher for the students taking the printed version of the test and the authors proposed that this may be due to the fact that the computer-based version of the text had PDF format. Here the test takers had to scroll and click to be able to read the full text which the authors suggested require a higher level of cognitive load of the test takers whereas the students taking the paper version could leaf through the test and keep it in their hands. However, it was not verifiable to what extent navigation and scrolling had an impact on the lower scores for the computer-based version. Mangen *et al.* also emphasize the lack of spatiotemporal markers for longer and more extensive texts when reading on a screen compared to a paper where the reader can tactilely and physically feel and see the dimension of the text. The result supported their hypothesis based on prior research by other scholars that reading comprehension on paper would gain better performance, although their explanation needs further research to be confirmed.

Another interesting find was that the computer-based group of test takers perceived that they had a better outcome compared to what they would have had on paper, and also preferred the digital texts to the printed equivalent.

Porion *et al.* (2016) examined 72 children' performance on paper-based and computer-based L1 (first language) reading comprehension and memorization in third and fourth year of secondary school. Prior to the computer-based test the test takers took a questionnaire to gain knowledge of their computer familiarity. The reading comprehension was a text about the bloodstream. The memory test consisted of 20 content words and the test takers were asked if they recalled the words from the text or if they were familiar with them before reading the text previously. There was no significant difference between the two modes. This did not go along with the prediction that reading from a computer would gain higher performances than reading from a screen. Their conclusion was that if all presentation conditions on screen and on paper (e.g. text structure, length, images, font type) are similar, reading performances on screen can become comparable to paper-based tests.

Singer & Alexander (2017) explored the reading comprehension of different texts on a computer or in print for 90 undergraduates. The participants were all part of a university course and not randomly selected. The topic of the texts in the study was childhood health and four book excerpts and four newspaper articles on the topic were presented. Surveys were administered prior to the test on topic knowledge, believed medium preference and a demographic survey. All test takers took the tests on print and on the computer. Singer & Alexander found that the recollection on key points in the texts was better when reading printed texts compared to those delivered on a computer (Mean score – main idea 2.64/2.56; Key points 5.61/5.19; other information 7.12/6.42). However, they found no difference in the outcomes for the two modes. Also this study reported that 69% of the test takers perceived that they performed better on the computer and 13% on paper.

An L2 (second language) writing assessment for students at Trinity College was investigated by Brunfaut *et al.* (2018). They examined the effect of delivery mode on two different writing tasks on the CEFR-levels B1-B2-C1. One test on paper and the other on the computer were taken on the same day by the test takers. There was also a questionnaire on the test takers' perception and some biographical data. 282 second language learners of English participated in the study. Their main purpose was to gain knowledge whether the paper-based tests were transferable to an online-format. Brunfaut *et al.* found no statistically significant difference in mean measures between the two modes. However, the test takers found it easier to edit and revise their texts on the computer-based version ($p < .001$). The test takers also perceived that they performed better on the computer-based test compared to the paper-based.

Writing assessment of English Foreign Language (EFL) on paper and computer was also investigated by Endres (2012). 28 Spanish students participated in the study. Prior to the test they completed a questionnaire to determine their computer skills. Two similar writing tasks were chosen to be delivered on paper and on the computer. The participants were divided into two groups, one group took the paper-based test first and then the computer-based test and the second group did the two tasks in the reverse order. Endres found little difference between the two delivery modes and no significant difference in mean achievement scores between the two delivery modes (PB 9,339; CB 9,679). The study showed however a statistically significant difference when looking at the higher mean for the number of words for the computer-based test (*t*-test indicated that 27.35 more words were used on average when taking the test on a computer). There was no statistically significant difference for word error between the modes. On the computer-based test, Endres found a large amount of orthographic errors caused by typing errors on the keyboard which had no similarities in the paper-based test. The test takers expressed a preference to taking the exam on a computer.

Jeong (2012) investigated the comparability of scores on paper- and computer-based test on 73 randomly selected Korean six-grade students. The test measured five different subjects taught in school: language arts, mathematics, social studies and science. In both delivery modes, the tests had 80 questions which were all multiple choice items. All participants took the paper-based test first and then the computer-based test. Prior to the computer-based test, the participants took part in training on how to navigate and interact on the computer-based test. The participants scored higher on the paper-based version of the test. However, ANOVA show that there was a significant difference only in Korean (p < .001) and science (p < .005). In this study, Jeong compared test scores for males and females both within groups and between groups. The findings indicate that all students performed better on the paper-based test and that the computer-based test score were significantly lower than the paper-based test scores for female students. Thus, male students achieved higher scores compared to female students. When investigating whether the average scores for males differed between the paper-based and computer-based test in the different subjects, a one-way ANOVA showed that there was a significant difference in only Korean. The same ANOVA was run for female students and the results show that there was significant difference in three of the subjects tested: Korean, Mathematics, and science. A conclusion drawn by Jeong, was that familiarity with ICT is higher for males and thus plays an important role on the scores. Based on the study, Jeong came to the conclusion that male students have an advantage since they spend more time on a computer by e.g. playing games compared to the female students. The article also concludes that the poor results of computer-based reading tests of Korean is due to the fact that reading on screen is harder than reading on paper. However, these interface issues will, according to Jeong, diminish over time as technology finds better ways of displaying large parts of texts on screen. Jeong suggests further studies into how different item formats effect the score in paper- and computer-based tests.

High computer familiarity was also found to facilitate the performances by the test takers in a study by Jin & Yan (2017). They examined 125 participants of a high-stakes English writing assessment in China. The participants were divided into four groups according to their English reading skills. The first two groups started with the paper-based test and then the computer-based test, and the third and fourth group took the tests in reverse order. A questionnaire on computer familiarity was administered after the computer-based test. All hand-written texts were entered verbatim into the computer by a typist. Jin & Yan saw that computer-based texts were assessed higher than paper-based texts for test takers with moderate and high computer familiarity. The higher scores were achieved for those with higher computer familiarity (p=.045). The test takers with low familiarity achieved a lower score on the computer-based version. The standard deviation for the computer-based test was higher than that of the paper-based test. The analysis also showed that the texts in the computer-based version were significantly lengthier and had fewer errors. This was also the case for the length of the sentences (Number of words, p <.006), with a large effect size (d= .84). Their main conclusion is however that computer- and paper-based writing tests gain similar scores. Nevertheless, they point out that it should be a priority to refine the construct of what is being measured.

Retnawati (2015) investigated whether the test results from an English proficiency test (L2) were equivalent on paper and on computer delivery. The data in the study came from randomly selected Indonesian test takers on paper-based- and computer-based tests. In each group there were 600 test takers. The computer-based version had almost the same reliability coefficients between the scores with classical test theory. With item response theory, the students with high and low ability had more accurate scores. For students with moderate ability the computer-based test was more accurate than the paper-based test. However, students unfamiliar with computers had problems logging in, using the mouse and similar technological problems.

Iranian university students were part of the study done by Hosseini *et al*. (2014) who investigated multiple choice items of L2 (English) reading comprehension of paper-based and computer-based tests. 106 randomly selected participants took two equivalent tests on paper and on the computer. First, they took the paper-based test and two weeks later the computer-based test. This was followed by a questionnaire into their computer familiarity and perception of the test. 10 randomly selected participants were also interviewed. Results showed that students preferred the computer-based version. However, computer familiarity and attitude towards the computer had no significant influence on the performance of the computer-based test (p > .014). The main conclusion was that the students performed slightly better on the paper-based version (mean 24.16/23.16). A *t*-test show significant difference (*p* < .001).

This conclusion was not supported by Khoshsima *et al.* (2017) when they looked at score comparability of a L2 English vocabulary test for 30 Iranian university students. These 30 students formed a homogenous group out of the 100 students who made out the total sample. The participants took the paper-based test first followed by the computer-based test. Their results showed no significant difference between the scores for the two delivery modes in the *t*-test. The mean was higher for the paper-based test by 0.53 points. Also, no difference between scores for male and female students was found.

Wang *et al*. (2008) conducted meta-analyses on 312 studies comparing assessments on computer or on paper for students from Kindergarten to 12th grade. The studies were all conducted between 1980 and 2005 and investigated reading achievement and ability. Wang *et al.* investigated the test scores from the studies. The meta-analyses of 36 homogeneous studies showed no statistically significant difference between paper-based and computer-based assessments. Their conclusion was that there is no difference between the two modes but they also stress that comparability studies will be needed in the future.

Barkaoui & Knouzi (2018) conducted an exploratory study on the effects of writing mode and computer ability on the scores of essays in an L2 writing test in Canada. The task was a TOEFL-iBT independent writing test. They also investigated the linguistic characteristics of the essays in the two modes. 184 essays written by 97 ESL (English second language). The students were divided into two groups depending on high and low ELP (English Language Proficiency) levels and also high and low keyboarding skills. The data analysis was based on eight factors: fluency, linguistic accuracy, syntactic complexity, lexical complexity, cohesion, text organisation, and content. The number of words per essay was a measure of fluency. The number of errors in a text was used as a measure of linguistic accuracy. Grammatical variation, mean length of T-unit, mean number of clauses per T-unit (sentence level), and syntactic variety were investigated to gain knowledges of the syntactic complexity. To assess lexical complexity they looked at range and size of the vocabulary of an essay. The number of connectives such as casual, additive, and temporal was an indicator of cohesion. The text organisation was measured by number and length of paragraphs. To gain knowledge of the content, the idea density and topic adherence were investigated. Barkaoui & Knouzi compared the words used in the prompts and task instruction with the essay and got an overlap of synonyms and words. They found that the participants wrote longer texts on the computer compared to those writing on paper (M =345.79 words for computer-based, and 277.34 for PB). Higher keyboarding skills lead to lengthier texts and higher scores (CB *r* =0.46, *p* < .01; PB *r* = 0.38, *p*< .05). However, they conclude that writing mode had no significant effect on any of the eight factors included in the analysis, only on writing scores.

Karay *et al*. (2015) investigated whether delivery mode influenced the test performance of 266 German medicine students. The test had 200 interdisciplinary multiple choice questions. Both tests were identical and students found either a paper-based test or a computer-based test on their table. The findings show that delivery mode had no impact on performance and the mean score did not differ significantly. However, in the computer-based version of the test students responded in significantly less time. They also found that low-performance students guessed more frequently on the computer-based version.

Maguire *et al.* (2010) conducted a study to investigate the difference in test scores between paper-based and computer-based tests for 179 accounting students. 43 of them took the test on the computer and 92 the paper-based version during regular class time. Both groups were given the same questions. They came to the conclusion that the scores was significantly higher for the computer-based test (mean 69.77; 64.177). A *t*-test show that there is a significant difference between the formats (*p* < .000164).

This was contradicted by Anakwe (2010) who looked at score comparability between paper-based and computer-based tests in accounting. 75 participants in an undergraduate accounting course at university took part in the study. They were administered four tests, two on paper and two on a computer. The first were paper-based test, followed by a computer-based test the second group in the reverse order order. The items were all multiple choice. *T*-test showed no significant difference across modes.

To address research in the area of equivalence between computer-based and paper-based testing and to contrast these with the findings made by Dillon in 1992, Noyes & Garland (2008) studied 19 articles published before 1992, thus before Dillon’s study, and 19 articles post 1992. The pre-1992 studies focused on traditional outcome measures such as reading speed, accuracy and comprehension and the studies favoured paper-based tests. Whereas the post-1992 studies focused on the complete tasks rather than the partial performance indicators mentioned above. A large number of these studies found no significant difference between the two modes. The authors also carried out a literature review of 41 articles from 1992 to 2005 concerned with online assessment using standardised tests. More than half of the studies reported no significant difference between computer-based or paper-based test. Some of the studies reported that the results from their tests were not equivalent because the test had been conceived and administered on paper and transferred to a computer-based test. But the psychometric properties of these tests were well-established and could consequently be checked against the computer-based version. The studies prior to the year 2000 reported anxiety issues from the test takers. Several of the studies reported test takers’ preference towards the computer-based test. Noyes & Garland summarized that future technology will probably achieve greater equivalence between modes.

As a conclusion, when considering the results from the articles included in this literature review, it is obvious that test takers prefer taking tests on a computer instead of on paper (Mangen *et al.* 2013; Singer & Alexander 2017; Brunfaut *et al*. 2018; Endres 2012; Hosseini *et al*. 2014). Several of the articles also concluded that test takers produce lengthier texts while writing on a computer (Jin & Yan 2017; Endres 2012; Barkaoui & Knouzi 2018). Something which was also noteworthy was the fact that males performed relatively better on computer-based tests according to Jeong (2012).

This research review indicated no consistent result with respect to whether the mode of test delivery matters for the students’ achievement. The single studies above did not find any differences in actual outcome of achievement performance between the two delivery modes. However, these studies are not only few and limited in scope, they are also in general small and suffer from weak design. The more recent of the two meta-analyses on the other hand, indicates the opposite, that the test delivery mode actually affects test performance. Thus, much more research is needed before any firm conclusions can be drawn. The reviewed studies have however reported other interesting findings in relation to computer-based testing, and also pointed out additional topics for future research.

4.3 Aim of this thesis

In the light of the theoretical background presented in the previous chapter and the finding from the literature review, this thesis aims to investigate the comparability of scores as well as the comparability of constructive responses when high-stakes English tests change delivery mode from paper-based to computer-based. Thus, the reliability of the tasks will be investigated to see whether the scores are comparable. A change in delivery mode can introduce bias or unfairness if certain groups are favoured if these have higher computer familiarity and digital literacy compared to other groups. This is irrelevant to the construct and will affect both reliability and validity of a test score. Are confounding variables inferred instead with a new delivery mode? Is there a difference in the test taker’s language performance caused by delivery mode? Is the same item comparable across two different delivery modes, paper-based or computer-based? There are many questions that arise on basis of previous research in addition to the theoretical background of language testing and measurement theory.

With these matters in mind, the present study was aimed to investigate the comparability of items delivered on paper and on screen and collect evidence of the influence of delivery mode on the different tasks. The empirical part of this paper is focused on a set of typical reading comprehension tasks in a test for year 9.

4.4 Research questions

The aims of the present study were broken down into the following research questions (RQ):

RQ1. Do the mean score and variance change depending on delivery mode?

Does reliability change depending on delivery mode?

Is item difficulty equal between delivery modes?

RQ2. How will the delivery mode affect the constructive responses given by the test takers?

RQ3. Is the test-taker’s perception of the task affected by the delivery mode?

RQ4. Are there different differences between boys and girls in the two delivery modes?

The general hypothesis, the null-hypothesis, is that the delivery mode will neither affect the performance of the test takers, nor the reliability of the test score.

5 Data, design and methods

All data in this study have been received from the large-scale pilot tests conducted within the NAFS-project as described in chapter 2. The selected test tasks have been piloted in both paper-based and computer-based tests in different schools and classrooms in such way that comparisons are possible in line with a quasi-experimental design, with the possibility to control for initial differences between the two groups with respect to English achievement as indicated by school grades in English.

5.1 About the data

Within the NAFS-project schools are regularly sampled to participate in similar tests. Access to schools are given via a school register delivered by the National Agency for Education. Schools are randomly selected to participate in the field trial. Selected schools receive a letter of invitation in which they are offered to choose which mode to participate in, either paper-based or computer-based or not participate at all. Schools without one computer per student, a lacking digital infrastructure such as high speed internet connection or the possibilities for professional support were unable to participate in the computer-based version. This self-selection procedure might have led to a bias between the groups and thus no random selection into either paper-based or computer-based test of the participating schools took place. To compensate for this lack of random assignment, information on other variables was collected so that selections effects could be cancelled out from the analysis. These variables were students' grades in English the previous term and students' results on some anchor tasks (this will be explained further in 5.3.2).

The participating schools had a span of three weeks to take the test. The computer-based version also included a demo test to take prior to the actual test in order to get acquainted with the platform and the process of logging in. The administration of paper-based tests followed well established standardised procedures and these are documented in the written instructions to the participating schools.

The data used in the current study was collected during the years 2016-2018 from large-scale pilot studies for future English high-stakes test for year 9 in compulsory school.

5.2 The participants

The participants (N= 1275) were 9th year students in Sweden which in this study are grouped into four groups according to what test and delivery mode they have participated in (see table 5.1). Group 1 consists of 358 students in total, and they received an extensive fictional reading comprehension paper-based test in November 2016. Group 2, consists of 310 test takers who received a non-fictional reading comprehension paper-based test with three different texts. Group 3, consists of 209 students who received the same non-fictional reading comprehension test as Group 2, but administered on computers. Group 4 was administered the same extensive reading comprehension test as Group 1 but on computers. There was large number of missing participants in the computer-based group of 2018. Approximately 131 test takers were unable to finish or even start up the computer-based tests. This large number of missing can be explained by technical problems during the test session or absences. Technical problems were mainly caused by lack of staff able to support the teachers, security levels for downloading a lock-down browser, or technical issues during the test. The number of missing cases at the individual level can therefore be assumed to be at random with respect to English language proficiency.

Table 5.1 Frequencies, statistics, number of participants

	2016	2017	2018
	Extensive reading comprehension	Short non-fictional reading tasks	Extensive reading comprehension
Paper-based test			
N	358	310	
Missing	36	17	
Computer-based test			
N		209	228
Missing		**79**	**52**

As explained above, this is a quasi-experimental study in that the schools and students included in this study were not randomly assigned different delivery modes. This indicates that the groups not necessarily are equal and comparable before the tests were administered. The schools in participating might be high or low achieving schools if analyses at school level had been carried out, this is pointed out by Coleman (1988) who explains that a school culture could show in the result. However, it is possible to make the groups more comparable by controlling for some initial differences, thus both gender and grade have been added as controls in the analysis. The grades collected were the grades the student was awarded in English the previous term. In this sample, all groups but the April-18 group reported the grade they got when finishing year 8 the previous term. The grade the group in April-18 was the autumn term grade in December. Thus, this group has had six months more of English education.

In this analysis, an F is awarded the value 1, C is 4, and A is 6. In the sample for this paper, the mean grade reported for the test takers of the paper-based group in 2016 is 3.71 and in 2017 3.74. The groups taking the computer-based versions in 2017 and 2018 had a mean grade of 3.85 and 3.90 respectively (see table 5.2). Depending whether the teacher supplied the requested information or not, there is a number of missing grades or gender in the study. These cases were omitted from the statistical analysis. The distribution of grades in the two format groups (PB or CB) is shown in table 5.2.

Table 5.2 Grade distribution in test groups

	Group 1 PB-test -16	Group 2 PB-test -17	Group 3 CB-test -17	Group 4 CB-test -18
Percent with grade A	11.2	12.9	10.4	13.6
Percent with grade B	20.1	21.0	22.5	19.7
Percent with grade C	23.7	24.8	26.8	28.5
Percent with grade D	24.0	15.2	14.8	21.9
Percent with grade E	15.1	20.3	12.9	13.6
Percent with grade F	5.9	5.8	8.6	2.6
Mean grade	3.71	3.74	3.85	3.90
Mean grade boys	3.53	3.63	3.53	3.78
Mean grade girls	3.88	3.85	4.19	4.04

5.3 Instruments

Two testlets made out the instruments for this thesis. One testlet consisted of a narrative text based on a literary fictional novel for adolescents and the other testlet consisted of three different non-fictional texts. These two testlets were administered on paper or on a computer to different groups of test-takers. The group taking the test on paper received A4 leaflets, and the group taking the test on a computer entered a testing platform via his/her computer. Depending on their school, the test takers used different digital devices.

5.3.1 Reading comprehension tasks

The five tasks included in this study are all part of large-scale pilot tests administered by the NAFS-project. In this study the different tasks are divided in two testlets, A and B. the testlets were delivered at different times, and with different modes. In table 5.3 detailed information about which tasks, in which modes, and when they were collected is displayed. Note that all test takers also participated in a paper-based subtest with items that are used as anchors, these are further described below.

Table 5.3 Test takers participating in the different sections

	PB-test 2016	PB-test 2017	CB-test 2017	CB-test 2018
Group	1	2	3	4
Testlet	A	B	B	A
Anchor (PB-mode for all test takers)	X	X	X	X
Matching task		X	X	
Information seek task		X	X	
Gap task		X	X	
Extensive reading task	X			X

There is a total of 52 different scored cognitive items across the four sections (Matching task, Information seek task, Gap task, Extensive reading task): Selected response (SR) items, including 16 multiple choice (four-options) and 12 multiple matching items, 25 short answer (SA) items where the test taker was supposed to enter only one word, and 11 constructed response (CR) items demanding more extensive responses.

Due to the confidentiality of the tasks, the actual tasks cannot be disclosed on a more detailed level. However, similar tasks already used in high-stakes tests for year 9 (CEFR B1.1) is instead described in the following section. They are also available on the NAFS web page. [4] Figures of the tasks are displayed in *Appendix 3*.

5.3.2 Anchor task

The anchor text consists of 12 items which each is awarded 1 score point for correct responses. The anchor task is a text with about 250 words in total, and the items are so called one-word gap items. One word is missing in each of the 12 different non-combined items and the test taker should fill in one missing word in each gap.

The anchor task is used to facilitate comparisons across groups and over time (Erickson, 2017) and it will show the proportion correct and thus, the test takers' proficiency level of English. A test taker with full score will get a proportion correct of 1.0 (100%) and the test taker who managed to complete half will get 0.5 (50%). In this study, each participant took the anchor task on paper to enable detection of initial differences between the four quasi-experimental groups.

[4] See: https://nafs.gu.se/prov_engelska/exempel_provuppgifter

5.3.3 Matching task

The second task consists of a matching task (about 250 words) with 12 items where the test taker should combine words with 12 explanations. There are in total 20 words to choose from. The task had a maximum score of 12. There is a visual difference between the layouts from the paper-based version to the computer-based due to the structure of the digital platform, where construction of items is limited to certain options provided. See examples in *Appendix 3,* figures 5.1 and 5.2. The examples are used for English 5 at upper secondary school but similar tasks are used for year 9.

5.3.4 Information seek task

The third task is an information seek text (about 900 words) where the test taker should seek and find information to complete the 14 items with questions. The items response formats are both open ended and multiple matching format. The task had a maximum score of 17. In *Appendix 3* figures 5.3 and 5.4 are examples of information seek tasks depicted.

5.3.5 Gap task

The fourth task is a so called gap text (about 400 words) where 14 gaps are spread throughout the text and the one word which is missing should be entered in each gap. The task had a maximum score of 14. In *Appendix 3* are two examples of a similar tasks presented in figures 5.5 and 5.6.

5.3.6 Extensive reading task

The fifth task is a longer extensive reading comprehension consisting of 2,000 words including both text and questions. The item format is both multiple choice and open ended questions. The large-scale pilot test of the paper-based version was administered to the students in December 2016 and the computer-based in April 2018. Originally, there were 21 items and a total score of 24. However, in between this two pre-tests some questions have undergone a few minor alterations and are thus omitted from this study. Thus, the task had in this study a maximum score of 17.

In *Appendix 3,* figures 5.7 and 5.8, show examples of an extensive reading comprehension task. The first example is form a paper-based test, the second from a computer-based task. There are some visual differences in appearance. In the paper-based test a picture and information on the task is located on the left page with the text below. The test items for this part of the text is located on the right hand side of the paper. Each page has an A4 format. The following section of the text is shown on the next page together with the questions for this part. An extensive reading comprehensions task is divided on 4-5 spreads of two-A4 papers.

The same extensive reading comprehension task on a digital test platform looked like the example in *Appendix 3,* figure 5.8. To minimize scrolling for the test taker, the instruction and illustration were placed on a page before the actual reading comprehension. To read the text and complete the tasks, the test taker need to moved between the pages with the arrows back and forth.

5.3.7 Presentation on paper and on computer

As seen in the examples in *Appendix 3,* all items on paper are on white background whereas the computer-based items have a grey background. The design and layout on the computer-based tasks, such as fonts or the font size, were not possible to change in this testing platform nor was it possible to adjust spacing between lines something which is possible on the paper-based tasks.

In the paper-based version, each test taker was given with an A4-leaflet. Here the tasks were presented with an instruction at the top of the paper followed by the tasks. The leaflet ended with a questionnaire where test taker feedback was collected. The test takers could move freely back and forth while taking the test and could choose in which order to answer the questions or, if they wanted, not give a response.

In the computer-based version different devices and monitors were used depending on the school's choice of equipment. The test takers participating in this study took the tests on either Mac, PC, Chromebook or iPads. The log-in procedure required approximately five keystrokes before the test taker could reach the test. A lock-down browser blocked all other communication or programs while the test takers fulfilled the tests. Depending on the task format, the instructions for each task were presented either above each task or on a previous page. The latter was the case for the extensive reading task.

In the computer-based test, the test taker had to scroll to be able to read the whole text in some of the tasks and depending on device used by the test taker. The test takers read and entered their responses on the screen and could use back-and-forward arrows to freely navigate in the test or move between the items by clicking on the item number in the list at the bottom of the screen. No items were obligatory to answer and the test takers could choose if they wanted to ignore an item. Finally, the test takers pressed the button "Hand in my test" when they had finished. Prior to the actual test, all test takers took a demo test as a tutorial on how to use the platform and start up the lock-down-browser. After the test they completed a questionnaire via an external link.

5.3.8 Questionnaire

Test taker feedback (TTF) and teacher feedback (TF) is an important and vital part of test development at NAFS. The questionnaires are included in the paper-based tests whereas the test taker feedback questionnaire for the computer-based test in this study was administered via an external link to an online questionnaire. A shortfall of commented questionnaires was observed within the sample group for the computer-based for autumn 2017, 147 out of 310 commented on the test taker feedback. According to the teachers this was due to lack of time. Fewer missing test taker feedback questionnaires were noted in the computer-based test in April 2018 when the questionnaire was administered in connection to the test on the testing platform.

For each task the test takers were asked to fill in a questionnaire which comprised five statements. The test taker reacted by choosing a number on a five-point Likert scale thus showing how well they agreed with the statements (from 1 - "absolutely not" to 5 - "yes, absolutely"). A higher the number indicated a more positive test taker attitude. This was followed by open-ended questions for each task. The questionnaires were similar for both delivery modes. However, in the test taker feedback for the computer-based test there was an additional final part of the questionnaire focusing on the experience of the platform, the device used, technical issues, preference computer-based or paper-based, familiarity with ICT, the experienced fairness of the test, and an open-ended question where test takers could comment (December -17). The questionnaire for the test taker feedback is presented in *Appendix 2*.

5.4 Setting and data collection

The collection of data took place in different EFL classrooms in year 9 in Sweden. All participating teachers and test takers were given the same instructions for the paper-based and computer-based tests.

As previously mentioned, the empirical study was part of an ongoing project within the NAFS-project according to the following procedures (Erickson, 2017). The process of the study is shown in bullet points below:

- The schools are randomly selected from a school register containing all schools in Sweden.
- Written invitation letters to the selected schools.
- Schools responded to the invitations and could choose to participate in either the paper-based test or the computer-based test.
- Written instructions sent to the participating schools together with the material needed (log-in papers, assessment instruction, information about the digital platform etc.). Information about anonymity of the test-takers' identity was included.
- Execution of tests either on paper or on a computer during a span of three weeks in November and December, 2017 for the non-fictional test. The fictional paper-based test took place in November 2016 and the corresponding computer-based in April 2018.
- The test takers for the computer-based test first took a demo test to get familiar the digital platform.

- After completing the tests, there were questionnaires to test takers and teachers to gain knowledge on how they perceived the tasks and the level of difficulty.
- Collection of material, both paper and digital.
- Coding of responses. Separate coding of test-takers' open-ended responses by two trained professional scorers.
- Comparisons of the coded responses, a third trained professional scorer participated. This led to a list of correct, acceptable and incorrect test-takers' responses.
- Statistical analyses of performance.

5.4.1 Scoring and coding

The test takers' responses on the paper-based tests were entered into the computer by a typist and checked to prevent bias in interpretations concerning test takers' handwriting.

Two trained professional scorers then coded the responses independently. The coded lists were compared and where there were differences, a third professional coder was consulted.

The selected response items in the computer-based test were scored automatically in the digital platform. However, for all items with open-ended followed questions, the same procedure for as for the paper-based tests was followed.

5.5 Data Analyses

5.5.1 Tools

Classical test theory (CTT) was used for the analysis of the data, *i.e.* mean, standard deviation, inter-item correlations and Cronbach alpha as a measure of reliability. The software used for the analyses was IBM SPSS Statistics 24.

The analysis is based on Douglas' (2010) method for understanding the meaning and value of test scores. The main concepts are: proportion correct indicating the item difficulty, the mean, the standard deviation, the reliability, and the standard error of measurement. Furthermore, factors outside the test have also been considered: particularity correlation, the t-test, ANOVA, and also regression analysis.

The mean gives the average score. Here, all scores are added up and then divided by the number of test takers. To gain further insight into the mean, the standard deviation which is a measure of variance, indicates the average number of points the test takers' scores deviate from the mean score.

The difficulty level of an item, proportion correct (*p*), gives the proportion of individual test takers in a sample that pass the item (Kline, 2005). A high value indicates that the item is easy, for example 1.00 indicates that everybody succeeded on this item (100%) whereas .5 indicates that 50 % of the test takers passed the item. In testing, items with *p* levels of 1.00 or 0.00 are of little use as they do not capture any differences among the test takers. However, some very easy items and some rather difficult items are needed to capture true differences as they do not discriminate among the test takers.

The standard error of measurement indicates how many points the test taker's score can vary if the test would be given again. This is a measure which inform about the confidence intervals around the estimate.

The correlation coefficient informs about the strength in the relationships between two variables. The correlation coefficient is an estimate which informs on how much overlap there is between variables. The value of 1.0 would indicate total overlap and thus a perfect relationship between the performances on the two tests. According to Douglas (2010), the correlation coefficient is very useful to establish whether a new task is doing its intended job. To gain knowledge if the value in the correlation coefficient is too low, the answer is in the statistical significance and probability. Kline (2005) explains how the strength and direction of the linear relationship between two variables are described with the correlation coefficient. The values vary from -1.00 to +1.00. Higher values indicate either a strong negative of a strong positive relationship, whereas weaker values show a more moderate relationship. The value of 0.432 indicates the chance of this finding is 432 out of 1000 times. The p-value of the correlation estimate informs about the significance level. When the p-value is smaller than 0.05, then the probability for correlation coefficient being obtained by chance, is less than 5 %.

To test the idea that there is no difference between test takers' scores in the two delivery modes a t-test was used in which the null hypothesis postulates that there are no differences, and the alternative hypothesis postulates that there are differences. T-tests are performed to see if the null hypothesis can be rejected. A t-value greater than 1.96 or -1.96 indicates support for the alternative hypothesis.

$H1_0$ = there will be no significant difference between scores in the two delivery modes.

$$t = \frac{(\bar{x}1\ - \bar{x}2)}{\text{estimate of the standard error of the difference between two sample means}}$$

A t-test investigates whether average scores from different tests are significantly different. The t-test determines if the difference occurred by chance. To gain knowledge of what the value of t-test indicates, there are tables of *Critical Values of t* to use. This table gives evidence if the result happened by chance. If there are more than two groups to be compared, the t-test is replaced by an Analysis of Variance (ANOVA).

The ANOVA investigates whether the variation between groups is larger than that within groups. The *F*-ratio calculates the variation between the groups divided by the variation within the groups. If there is no variation between the groups, the larger the value the more probable it is that there is a difference between the groups and that this difference did not occur by chance. In order to establish if the *F* obtained is large enough to indicate that the results did not occur haphazardly, there are tables of critical values for ANOVA. The degree of freedom (df) has to do with the number of groups (take the number of groups minus 1).

The reliability of a test score indicates how trustworthy these results are. To calculate the reliability of a test score some the split-half methods are usually used. A test is split in two halves and then the correlation between the two halves is calculated. The higher the correlation the better (1.00 indicates total correlation).

According to Field (2018), the most frequent measure of scale reliability is Cronbach' alpha, α coefficient. It is used to investigate to what degree the test items in the test correlate, and the overlap indicates to what degree the items measure the same construct. Cronbach's alpha is a function referring to the number of items, the average covariance between item-pairs and the variance of the total score, and sample size. The Cronbach's alpha may vary between 0 and 1. An estimate above .7 should be considered good and sufficient at a group level, while values above .9 are targeted for reliable measures of individuals. Values above .7 can be hard to obtain when samples are small and/or the number of items in the test are few.

Linear regression analysis investigates the linear relationships between different variables, one dependent and the other/s independent and if any of them influence the test takers' performance on the computer-based test. The beta (β) value received in a linear regression analysis shows the relationship between the predictor and the criterion (Kline, 2005). The criterion is the dependent variable and the independent/s the predictor. This value ranges from -1.00 to 1.00. A score of 0.00 would indicate that there is no relationship between the predictor and the criterion which would be confirmation that the data met the prerequisites for the analysis (normality, homoscedasticity, linearity, no multicollinearity, no outliers).

Linear regression analysis can also be used to investigate group differences, with controls of covariates. In this study regression analysis techniques are used to investigate differences between delivery modes, while at the same time controlling for grade, sum of anchor task and gender. To investigate possible interaction effects two variables were multiplied together, hence an interaction element. Following Field (2018), grand mean was used for the investigation of possible interaction effects between gender and delivery mode.

To address the research question "*How will the delivery mode affect the constructive responses given by the test takers*?" half of the extensive constructive responses were selected for a word count. The decision to only investigate half of the constructive responses was due to the cumbersome task and that it would not be possible to administer within the time frame for this master thesis.

5.6 Ethical Considerations

According to the ethical principles of the Swedish Research Council (Vetenskapsrådet, 2017) participating schools and students received written information on the background, purpose and use of the study. All participants in the NAFS data base were anonymized and no other information except gender, grade and performance on the anchor tasks were used in this study. In large-scale pilot test, the test takers were only assigned a chronical case number without any links to the test takers' names, or school etc. Consequently, all test takers are anonymous.

5.7 Limitations

This thesis limits its scope to receptive reading tasks for year 9 of compulsory school, corresponding to the CEFR level B1.1. Accordingly, the focus is on high-stakes, large-scale English reading comprehension assessment. All items included in this paper were part of the on-going pre-testing of tasks and have been tested both on paper and on computer by different groups. To enable comparison across groups, each test taker has taken the same anchor item on paper. Thus, the English proficiency level of the groups in the study are comparable to previous pre-large-scale tests done by the NAFS-project. As the items tested within this study are part of future high-stakes tests, they cannot be displayed in this paper. However, similar items which are no longer restricted by secrecy and thus public, will be used to clarify the items in this study (see *Appendix 3*). All the different item formats used in a national test are not included in this study.[5]

As mentioned previously, the design resulted in a quasi-experimental study and the participating groups were not necessarily comparable something which made comparisons more difficult. The statistical analyses conducted were all based on classical test theory. Applying Item Response Theory, IRT, may have given additional information and knowledge on the comparability of scores on computer-based and paper-based tests, however such advanced techniques require much more time and knowledge and is therefore suggested for further research on these data.

[5] For more examples see: https://nafs.gu.se/prov_engelska/exempel_provuppgifter/engelska_ak9_exempeluppg

6 Results

In this chapter the results from the statistical analysis of the data collection are presented. To begin with, the Cronbach's Alpha values will give insight into the reliability of the tasks included in the study. Then, the focus will be to establish whether the different groups taking the tests are comparable. This will be followed by analysis into the total score of the two testlets. This is then followed by an analysis on item level for the different parts. To sum up, the information of the items will be presented with the test taker feedback collected on the different tasks. In a conclusion, a short summary of results will be presented.

6.1 Total score

As an attempt to answer research question "*Does the reliability of the test result change depending on delivery mode*?" reliability measures were investigated. Testlet A, the extensive reading comprehension task, had a Cronbach's alpha of .83 for the paper-based test and .86 for the computer-based test. Testlet B, the three shorter non-fictional tasks, received a Cronbach's Alpha of .90 for the paper-based test and .89 for the computer-based test.

The research question "*Does the mean score and variance change depending on delivery mode*?" investigated whether mode plays any part on the scores. In table 6.1 the descriptive statistics of the two testlets are displayed. The mean score of the two delivery versions show two different results. For testlet A, the extensive reading task, the mean was slightly higher for the computer-based test, 11.76 and 11.50 for the paper-based test. For testlet B, there was a higher mean score for the computer-based mode. The mean for the computer-based test was 27.28 compared to the paper-based test which scored a mean of 24.60 (see Table 6.1).

Table 6.1 Descriptive statistics, scores on testlets A and B

	N	**Mean**	**Median**	**Std. Error**	**Std .Dev.**	**Variance**
Testlet A, PB mode	330	11.50	12	.214	4.13	17.08
boys	173	11.24		.304	4.00	15.94
girls	157	11.81		.341	4.27	18.24
Testlet A, CB mode	228	11.76	13	.283	4.48	20.05
boys	120	11.38		.430	4.66	21.72
girls	108	12.19		.409	4.25	18.04
Testlet B, PB mode	310	24.60	26	.675	11.89	141.28
boys	163	25.90		.963	12.30	151.28
girls	147	23.20		.930	11.28	127.24
Testlet B, CB mode	181	27.28	29	.643	10.53	110.80
boys	92	26.32		1.11	10.70	114.31
girls	**89**	**28.3**0		**1.10**	**10.32**	**106.45**

The first step in the analysis was to identify whether there were initial differences between the groups taking the paper-based tests and the computer-based tests as the groups are not randomized. An independent t-test was conducted to compare the grades for the groups in the two delivery modes. There was no significant difference in mean grade for the groups taking the paper-based or the computer-based tests (Testlet A: M = 3.71 PB, SD 1.39, 3.90 CB, SD 1.32; *t* (584) =-1.67, *P* = .10, two-tailed; Testlet B: M =3.74 PB, SD 1.47; 3.85 CB, SD 1.5; *t* (517) = -.843, *P* =.705). When comparing grades for each gender, the girls have a higher mean grade in all sample groups (see table 5.2). An independent t-test was conducted to compare the grades for each gender in the two delivery modes. There was no significant difference for the grades for boys (*P* = .164 Testlet A; *P* = .383 Testlet B) in test groups who took the extensive reading comprehension test (testlet A) and the shorter mixed reading comprehension test (Testlet B). For the girls there was significant difference in the extensive reading comprehension test (Testlet A) (*P* = .001) but not in testlet B (*P* = .827). There are differences which need to be explored further and to investigate whether these differences are significant regression analyses were carried out.

A one-way ANOVA analysis is used to investigate whether there is statistically significant difference between the means for boys and girls sowed with delivery mode as factor showed non-significance for gender or mode for testlet A. For testlet B there was non-significance for gender (p=.924) but a significant difference for mode effect on total sum of testlet B (p=.012).

To address the research question "*Are there different differences between boys and girls in the two delivery modes?"* the mean scores were investigated which is displayed in table 6.1. A closer look at the mean score for both genders show that the girls of testlet A scored higher in both delivery modes, in the paper-based test 11.24 for boys and 11.81 for girls and the computer-based test show 11.38 for boys and 12.19 for girls. Testlet B has higher mean for boys in paper mode (25.9 boys and 23.2 girls) and the computer-based test have a higher mean for the girls (26.3 boys and 28.3 girls). The number of test takers on the paper-based testlet A differs from total number of participants (330 out of 358) and on the computer-based testlet B (181 out of 209). This difference is due to test takers not completing their test or absences on the day of the test.

- Testlet A

To address the research questions, a series of regression analyses were carried out. These analyses are presented below. The test score for testlet A, which was the dependent variables, centered grade and mode were introduced in the first model as an independent variable (see table 6.2). The results revealed no statistically significant difference for the variable mode between paper-based and computer-based tests (P=.850) but statistically significant difference for grade (P=.000). The R^2 value showed that the model predicted 47.1% of the scores on testlet A (R=.686; R^2 .471).

Table 6.2 Sequential regression with reading comprehension scores (Testlet A – extensive reading comprehension task) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
1	Intercept Score testlet A	11.500	.235		48.846	.000
	Variable – centered grade	2.188	.099	.686	22.202	.000
	Variable – mode (1=CB-test)	-.051	.269	-.006	-.190	.850

Dependent variable: Total sum testlet A

* *p*-Value is considered statistically significant at *p* <.05.

In table 6.3 the sum of the anchor task was used in the regression analysis instead of grade as shown in the table above (table 6.2). When controlling for anchor task, the *p*-value was significant on a 10 % level. The items in the anchor task could have a stronger correlation with the other tasks included in the tests as they more similar than the variable grade and tend to measure the same construct. As the scope of the anchor is not as broad as that of the grade the following regression analyses will use grade as variable instead of the variable anchor sum.

Table 6.3 Sequential regression with reading comprehension scores (Testlet A – extensive reading comprehension task) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
1	Intercept Score testlet A	2.205	.466		4.735	.000
	Variable – anchor sum	1.085	.050	.693	21.573	.000
	Variable – mode (1=CB-test)	-.461	.279	-.053	-1.654	.099

Dependent variable: Total sum testlet A

* *p*-Value is considered statistically significant at *p* <.05.

To be able to verify if there were gender differences, the regression analysis had gender as independent variable. The results reveal no significant difference for gender on the test score (see table 6.4). The following step was to add mode to the model. No significant difference was found.

Table 6.4 Sequential regression with reading comprehension scores (Testlet A – extensive reading comprehension task) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
1	Intercept Score testlet A	11.295	.250		45.253	.000
	Variable – centered grade	2.188	.099	.687	22.006	.000
	Variable – mode (1=CB-test)	-.060	.269	-.007	-.222	.824
	Variable – gender (0=boys; 1=girls)	-.046	.267	-.005	-.171	.864

Dependent variable: Total sum testlet A
* *p*-Value is considered statistically significant at *p* <.05.

Finally, the interaction element was entered in the regression. The output showed only statistically significant differences for the predictor grade. See table 6.5.

Table 6.5 Sequential regression with reading comprehension scores (Testlet A– extensive reading comprehension task) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
1	Intercept Score testlet A	11.673	.239		48.915	.000
	Variable – centered grade	2.190	.099	.688	22.015	.000
	Variable – mode (1=CB-test)	-.287	.371	-.033	-.773	.440
2	Variable – gender (0 boys/1 girls)	-.242	.346	-.028	-.699	.485
3	Variable – interaction element (mode*gender)	.478	.538	.044	.889	.374

Dependent variable: Total sum testlet A
* *p*-Value is considered statistically significant at *p* <.05.

The constant shows the scores of boys on paper-based tests. To calculate the scores of boys on the computer-based test the β coefficient for the variable mode is non-significant. Thus, there are no differences on scores for boys on either delivery mode (0.5 points higher mean if the test was taken on a computer).

- Testlet B

The same analysis was carried out for testlet B. Using the Stepwise method, a multiple regression analysis with the scores of the testlet as a dependent variable was performed. The regression analysis showed a p-value for the variable mode which indicates no statistically significant difference but statistically significant difference was found for grade (P=.000). See table 6.6 below. The R^2 value showed that the model explained 54.4% of the scores for testlet B (R= .738; R^2 .544).

Table 6.6 Sequential regression with reading comprehension scores (Testlet B) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
3	Intercept Score testlet B	24.859	.441		56.392	.000
	Variable – centered grade	5.821	.244	.733	23.842	.000
	Variable – mode (1=CB-test)	.885	.730	.037	1.212	.226

Dependent variable: Total sum testlet B
* *p*-Value is considered statistically significant at *p* <.05.

The next step in the regression analysis, was to add gender in the same regression (see table 6.7). The results reveal that mode does not play any significant role in the outcome. Thus, the null hypothesis, that there will be no difference between the two modes, is supported. Significant relations (P= .000) are indicated for gender and grade. In total, boys would thus almost increase their score with 3 points if they had the same grade but took the test on a computer. The R^2 value showed that the model explained 56% of the scores for testlet B (R= .750; R^2 .560).

Table 6.7 Sequential regression with reading comprehension scores (Testlet B) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
4	Intercept Score testlet B	26.348	.544		48.449	.000
	Variable – centered grade	5.957	.241	.750	24.683	.000
	Variable – mode (1=CB-test)	.898	.716	.038	1.254	.210
	Variable – gender (0 boys/1 girls)	-3.126	.693	-.136	-4.512	.000

Dependent variable: Total sum testlet B

* *p*-Value is considered statistically significant at *p* <.05.

In order to investigate whether the scores for girls and boys differ due to delivery mode, an element of interaction was added to the regression (see table 6.8). The analysis show that considering the format there is an effect of interaction. The constant shows the scores of boys on paper-based tests. Results for gender show that boys have 4 points higher score on the paper-based (girls scored: 26.768 – 4.015 = 22.753). However, by controlling for grade the interaction element becomes non-significant indicating that with the same grade for both genders there are no differences in performance depending on delivery mode. These findings are in agreement with previous research and results on national tests. The question has to be addressed if there are further benefits depending on gender and delivery mode.

Table 6.8 Sequential regression with reading comprehension scores (Testlet B) as the main outcome

Model	Source	*B*	*SE B*	β	*t*	*P**
5	Intercept Score testlet B	26.768	.596		44.917	.000
	Variable – centered grade	5.931	.241	.747	24.570	.000
	Variable – mode (1=CB-test)	-.275	.991	-.012	-.277	.782
	Variable – gender (0 boys/1 girls)	-4.015	.865	-.175	-4.639	.000
	Variable – interaction element (Mode*gender)	2.432	1.425	.082	1.707	.088

Dependent variable: Total sum testlet B

* *p*-Value is considered statistically significant at *p* <.05.

6.2 Task and item analysis

To be able to answer research questions "*Is item difficulty equal between delivery modes?*" the following section will investigate each task and the items included. The next section looks closer at the anchor task, this section will be followed by the matching task, the information seek task, the gap task, and the extensive reading task. Finally, the section will finish with quotes from the test taker feedback.

6.2.1 Anchor task

The anchor task had a total score of 12 and was administered on paper. The anchor task facilitates comparisons across time and groups and for this reason the task was administered on paper to both groups of test takers (PB and CB). The Cronbach's alpha for the paper-based test is .87 (PB-test -17), .79 (CB-test -17), .84 (PB-test -16) and .82 (CB-test -18). The proportion correct gives an indication of the difficulty of the item, as for dichotomously scored items (scores either 0 or 1), the mean equal of the proportion correct responses was .95 which means that 95% of the respondents have received score 1 on this item. This value is also an indication of the language ability of the group, meaning the proportion of students with all items correct. The group taking the paper-based test in 2016 had an item mean of

.69, the paper-based group 2017 scored .77, the computer-based group 2017 got .71, and the computer-based group in 2018 had a value of .75. The mean score for the groups taking the paper-based test was 8.45 and the computer-based group had a mean score of 9.20 (see table 6.9). The anchor task was not included as a variable in the regression analysis, except for testlet A the first model, as the scope it was not a broad as the variable grade and it showed a roof effect.

Table 6.9 Descriptive statistics anchor task, all test takers taking the task on paper

	N	Mean	Standard Error of Mean	Median	Std. dev.	Variance
Paper-based test	**617**	**8.45**	**.120**	**10**	**3.08**	**9.48**
Boys	322	8.57			3.06	
Girls	295	8.24			3.09	
PB- testlet A - boys	171	8.45			2.92	
PB testlet A - girls	159	8.36			3.01	
PB testlet B - boys	151	8.71			3.21	
PB testlet B - girls	136	8.10			3.19	
Computer-based test	**448**	**9.20**	**.115**	**10**	**2.60**	**6.78**
Boys	205	9.27			2.45	
Girls	188	9.19			2.78	
CB- testlet A - boys	110	9.46			2.50	
CB testlet A - girls	106	9.13			2.66	
CB testlet B - boys	95	9.06			2.40	
CB testlet B - girls	**82**	**9.27**			**2.93**	

The anchor task was administered on paper to all test-takers to enable comparison between the groups. As is displayed in the table above (6.10), boys scored a higher mean compared to the girls (e.g. M =8.57 PB, 8.24 CB). The group with the highest mean score on the anchor task was the group taking the computer-based extensive reading comprehension task (M = 9.46 for boys and 9.13 for girls).

A closer look at each of the 12 items show that the group taking the computer-based test in December 2017 performed better than the three other groups. Their proportion correct is .77. The computer-based test group in April 2018 also scored a higher proportion correct level compared to the paper-based pre-tests (see Table 6.10).

Table 6.10 Proportion correct, anchor task. Boys and girls within test delivery mode

Section1 - Anchor	PB-group -17	CB-group -17	PB-group -16	CB-group -18
Item 1	**.95**	**.95**	**.95**	**.95**
Boys	.96	.94	.95	.94
Girls	.95	.96	.95	.96
Item 2	**.57**	**.69**	**.62**	**.71**
Boys	.66	.72	.66	.72
Girls	.57	.70	.57	.70
Item 3	**.78**	**.87**	**.78**	**.87**
Boys	.78	.86	.78	.86
Girls	.77	.88	.77	.88
Item 4	**.79**	**.78**	**.78**	**.78**
Boys	.79	.77	.79	.77
Girls	.77	.79	.77	.79
Item 5	**.70**	**.66**	**.72**	**.62**
Boys	.75	.56	.75	.56
Girls	.68	.70	.68	.70

Item 6	**.82**	**.90**	**.85**	**.90**
Boys	.87	.93	.87	.93
Girls	.83	.88	.83	.88
Item 7	**.64**	**.79**	**.68**	**.79**
Boys	.72	.79	.72	.79
Girls	.63	.79	.63	.79
Item 8	**.83**	**.87**	**.80**	**.85**
Boys	.83	.83	.83	.83
Girls	.77	.88	.77	.88
Item 9	**.82**	**.85**	**.76**	**.85**
Boys	.78	.86	.78	.86
Girls	.74	.83	.74	.83
Item 10	**.18**	**.25**	**.19**	**.24**
Boys	.19	.21	.19	.21
Girls	.20	.28	.17	.28
Item 11	**.51**	**.68**	**.53**	**.72**
Boys	.60	.71	.60	.71
Girls	.45	.73	.45	.73
Item 12	**.75**	**.83/**	**.76**	**.8**
Boys	.78	.88	.77	.88
Girls	.73	.85	.73	.85

6.2.2 Matching task – testlet B

The matching task had a total score of 12. The reliability statistics give a Cronbach's alpha value of .89 (PB), and .84 (CB). The proportion correct was .56 (PB) and .64 (CB). In table 6.12 the statistics for the matching task is displayed. The mean was higher for the computer-based test (M 6.75 and 7.74) and the standard deviation was lower (3.31 to 3.84).

Table 6.11 Statistics, Matching task

	N	Mean	Standard Error of Mean	Median	Std. dev.	Variance
Paper-based test	306	6.748	.22	7.00	3.84	14.77
Computer-based test	178	7.741	.25	9.00	3.31	10.97

A closer look at each of the 12 items gives the following result (see Table 6.12). A one-way ANOVA between the two delivery modes show a statistically significant difference on 5 of the 12 items (2, 3, 4, 8, and 9). When investigating the items closer, no immediate explanations were found as to why these items got a higher proportion correct on the computer-based test. In all items except for item 5, the proportion correct is higher for the computer-based mode. All computer-based items scored a higher mean compared to the paper version.

Table 6.12 Proportion correct of task 2, matching

Item	Proportion correct PB-test	Proportion correct CB-test	Sig.*
1	.58	.67	.052
2	.63	.78	.000
3	.68	.82	.001
4	.82	.90	.009
5	.61	.61	.922
6	.50	.56	.168

7	.35	.40	.280
8	.36	.52	.001
9	.62	.72	.034
10	.64	.71	.150
11	.53	.59	.175
12	.43	.46	.664

** Sig. value statistically significant (p < .005).*

The test taker feedback shows that the test takers were generally positive towards both delivery modes with a mean of 3.44 for the paper-based version when commenting the statement “It was a good test” and higher mean for the computer-based test of 3.80 on a Likert-scale of 5. The statement “I think I did well” got a mean of 3.81 for the computer-based version compared to 3.23 for the paper-based.

6.2.3 Information seek task – testlet B

The information seek task consisted of 14 items and with a total score of 17. The Cronbach’s alpha for the paper-based test is .86 and the proportion correct indicating the item difficulty of .70. For the computer-based test, the alpha value is .80 and proportion correct .78. The mean was higher for the computer-based test (M 9.89 and 11.60). The standard deviation of the item, the paper-based version has a value of 4.80 and the computer-based 3.91 (see Table 6.13).

Table 6.13 Statistics, information seek

	N	Mean	Standard Error of Mean	Median	Std. dev.	Variance
Paper-based test	309	9.89	.27	11.00	4.80	23.08
Computer-based test	178	11.60	.29	12.00	3.91	15.26

A closer look at each of the 14 items gives the following result which is displayed in table 6.14. A one-way ANOVA comparing the mean score of the two modes show a statistically significant difference for 12 items out of 14. A comparison of each of the 14 items, shows that all but two (item 1 and 10), in the computer-based version scored a higher mean than those in the paper-based version. No obvious explanation was found to this difference when examining the items closer.

Table 6.14 Proportion correct, information seek

Item	Format	Proportion correct	Proportion correct	Sig.*
		PB-test	CB-test	
1	MC	.90	.88	.469
2	CR	.70	.79	.030
3	CR	.43	.50	.035
4	CR	.57	.66	.048
5	CR	.33	.40	.085
6	CR	.48	.58	.019
7	MC	.71	.81	.008
8	MC	.63	.78	.000
9	MC	.61	.78	.000
10	MC	.52	.41	.000
11	MC	.72	.86	.000
12	MC	.54	.58	.350
13	MC	.61	.69	.092
14	MC	.65	.71	.172

** Sig. value statistically significant (p < .005).*

The first item and items 7 to 14 are multiple choice items, whereas items 2 – 6 are open ended, constructed responses free to more extensive responses. However, two of these six items require only short answers. In order to get some knowledge in whether the delivery mode plays any role for the test takers results an investigation of three of the six open ended items was carried out. Item 2 with a proportion correct of .80 (CB-test) and .70 (PB-test), is a fairly easy item. Item 4 had .66 (CB) and .57 (PB) and is consequently a somewhat harder item. In both delivery modes, item 4 was the hardest of the three items.

To address the research question "*How will the delivery mode affect the constructive responses given by the test takers*?" half of the constructive responses that require a more extensive response for the test takers were selected for closer investigation. In the computer-based mode the average length of a response for item 2 was 10.3 words and for item 4 13.0. In the paper-based mode the values are 7.6 words for item 2 and 4.7 for item 4. The lengthiest response for item 2 in the paper-based mode consisted of 20 words. The mean number of words for item 5 is 35 for paper-based version and 37 for computer-based. When comparing the responses given by the test takers in the two modes, the computer-based test shows a higher number of words per item. These results are displayed in tables 6.15-6.17.

Table. 6.15 Analysis of item 2, information seek

Mode	Proportion correct	Mean number of words/response	Max number of words/response	Min number of words/response
Paper-based test	.70	7.6	20	1
Computer-based test	.80	10.35	40	1

Table. 6.16 Analysis of item 4, information seek

Mode	Proportion correct	Mean number of words/response	Max number of words/response	Min number of words/response
Paper-based test	.57	4.7	35	1
Computer-based test	*.66*	13.0	37	1

Table. 6.17 Analysis of item 5, information seek

Mode	Proportion correct	Mean number of words/response	Max number of words/response	Min number of words/response
Paper-based test	.32	7.0	24	1
Computer-based test	.40	12.0	44	1

The test taker feedback showed a more positive attitude to the computer-based test. The statement "It was a good test" had a higher mean for the computer-based version of 3.65 compared to 3.37 for the paper-based version. The statement "I think I did well" received a mean of 3.62 for the computer-based and for the paper-based 3.08.

6.2.4 Gap task – testlet B

309 test takers took the paper-based version and 169 the computer-based version of the one word gap task. The Cronbach's Alpha is .89 for the paper-based test and .84 for the computer-based. All the 14 items contributed to the Alpha value. The proportion correct for text on the paper-based test is .57 and for computer-based version .64. The mean score for the computer-based test was 8.85 and 8.10 for the paper-based test. This is the only task included in this study with such a low difference between the mean score of the two delivery modes (see Table 6.18).

Table 6.18 Statistics, gap task

	N	Mean	Standard Error of Mean	Median	Std. dev.	Variance
Paper-based test	309	8.10	.239	9.00	4.19	17.58
Computer-based test	169	8.85	.280	9.00	3.64	13.25

A closer look at the mean score for each item shows that all 14 items scored a higher mean on the computer-based test (see table 6.19). A one-way ANOVA between the two delivery modes show a statistically significant difference on 5 of the 14 items (3, 5, 7, 11, and 14). Only item 12, which is the most difficult item in this task, scored a higher mean on the paper-based test. No obvious explanation was found to this difference when examining the items closer. Otherwise, all other items got a higher mean on the computer-based test.

Table 6.19 Proportion correct, gap task

Item	Proportion correct PB-test	Proportion correct CB-test	Sig.*
1	.73	.80	.087
2	.51	.54	.490
3	.71	.79	.045
4	.49	.52	.462
5	.79	.89	.009
6	.57	.65	.083
7	.61	.72	.013
8	.43	.46	.513
9	.40	.45	.274
10	.69	.68	.900
11	.83	.91	.018
12	.40	.36	.388
13	.66	.68	.603
14	.29	.39	.033

** Sig. value statistically significant (p < .005).*

The test taker feedback shows that the test takers were generally positive. The statement “It was a good test” got the same mean for both modes, 3.84. The statement “I think I did well” got a mean of 3.73 for the computer-based test and for the paper-based version 3.43.

6.2.5 Extensive reading task – testlet A

The extensive reading comprehension task had a total score of 17. The reliability coefficient, Cronbach’s alpha for the paper-based test is .83 and .86 for the computer-based test. The proportion correct indicating the item difficulty for the paper-based test is .76 and .77 for the computer-based test. The mean was slightly higher for the computer-based test (M 15.90 and 16.08). What is also indicated in the standard deviation of the item, the paper-based test has a value of 5.32 and the computer-based test 6.33 (see Table 6.20).

Table 6.20 Statistics, extensive read comprehension task

	N	Mean	Standard Error of Mean	Median	Std. dev.	Variance
Paper-based test	362	15.90	.279	17.00	5.32	28.31
Computer-based test	280	16.08	.378	18.00	6.33	39.99

This task shows a different result compared to previous items. Comparisons of the proportion correct with the results from the anchor task indicate that boys and girls score more or less similar. In this extensive reading comprehension task, the boys scored a higher mean on the paper-based test on two items and in the computer-based test on five items, whereas in the anchor task, boys had a higher item mean on all items (CB-test -17), on 11 out of the 12 items (PB-test -17, CB-test -18), and on 8 out of 12 items (PB-test -16) (see table 6.21). When looking at the mean for each item, 8 out of 15 scored higher mean value on the paper-based version. Out of these 8 items, five were multiple choice items. On item 5, both modes scored the same mean. When comparing the two genders, a one-way ANOVA show no significant differences on the computer-based test, whereas there is a statistically significant difference between boys and girls on the paper-based test on item 2 (P= .002).

Table 6.21 Proportion correct, extensive reading comprehension task

Item	**Proportion correct**	**Proportion correct**	**Sig.***	**format**
	PB-test	**CB-test**		
1	**.80**	**.79**	**.899**	**MC**
Boys	.80	.80		
Girls	.80	.78		
2	**.51**	**.55**	**.280**	**CR**
Boys	.42	.56		
Girls	.60	.55		
3	**.53**	**.62**	**.033**	**MC**
Boys	.52	.60		
Girls	.54	.64		
5	.57	.67	.017	MC
Boys	.55	.67		
Girls	.59	.66		
7	**.78**	**.71**	**.059**	**CR**
Boys	.80	.72		
Girls	.76	.71		
10	**.77**	**.71**	**.196**	**MC**
Boys	.74	.72		
Girls	.80	.71		
11	**.63**	**.71**	**.045**	**MC**
Boys	.62	.67		
Girls	.64	.75		
12	**.83**	**.79**	**.151**	**CR**
Boys	.82	.74		
Girls	.85	.83		
14	**.68**	**.75**	**.076**	**CR**
Boys	.64	.70		
Girls	.71	.80		
15	**.47**	**.52**	**.267**	**CR**
Boys	.42	.47		
Girls	.52	.57		
16	**.62**	**.62**	**.995**	**MC**
Boys	.62	.57		
Girls	.62	.68		
17	**.77**	**.82**	**.158**	**CR**
Boys	.78	.78		
Girls	.77	.86		

18	**.64**	**.61**	**.543**	**MC**
Boys	.62	.58		
Girls	.66	.65		
19	**.79**	**.73**	**.153**	**CR**
Boys	.77	.73		
Girls	.78	.74		
20	**.59**	**.61**	**.658**	**MC**
Boys	.56	.56		
Girls	.63	.67		

** Sig. value statistically significant (p < .005).*

To address the research question "*How will the delivery mode affect the constructive responses given by the test takers*?" half of the constructive responses that requires a more extensive response for the test takers were selected for closer investigation. As shown in table 6.22, the lengthiest response in computer-based test had 66 words whereas in the paper-based version the corresponding value was 49. The mean length was 17.8 for the computer-based task and 13.3 for the paper-based.

Table 6.22 Analysis of item 5, extensive reading comprehension task

Mode	Proportion correct	Mean number of words/response	Max number of words/response	Min number of words/response
Paper-based test	.77	13.3	49	3
Computer-based test	.76	17.8	66	2

Item 14 had a higher proportion correct in the paper mode than in computer mode (.72 and .67). A word count showed that the mean length of answers in the paper-based task was 7.69 compared to 10.37 in the computer-based task. The longest response in the paper-based task had 29 words compared to 33 words in the computer-based version (see Table 6.23).

Table 6.23 Analysis of item 14, extensive reading comprehension task

Mode	Proportion correct	Mean number of words/response	Max number of words/response	Min number of words/response
Paper-based test	.67	7.69	33	3
Computer-based test	.72	10.37	29	2

As shown in table 6.24, item 15 had the same proportion correct for both modes, namely .49. A word count showed that there was barely any difference between computer-based (8.239) and paper-based (8.19). The longest response in the computer-based task had 34 words compared to 26 in paper-based task.

Table 6.24 Analysis of item 15, extensive reading comprehension task

Mode	Proportion correct	Mean number of words/response	Max number of words/response	Min number of words/response
Paper-based test	.49	8.19	26	1
Computer-based test	.49	8.24	34	1

The test taker feedback shows that the test takers were generally positive with a mean value of 3.86 for both modes on a Likert-scale of 5. However, those taking the paper-based test thought the text was more difficult compared to those taking the computer-based test (2.87 PB, 3.21 CB). When reflecting whether they thought they did well those taking test on paper were less positive 3.42 compared to those taking it on the computer 3.61.

6.2.6 Summary of all tasks

The result of the analyses based on classical test theory indicates that the reliability coefficients of the scores on the paper-based and the computer-based tests are high. Closer investigation of all items included and their reliability coefficients showed no major discrepancies.

6.3 Summary of test taker feedback

Test taker feedback has been reported along with the different tasks in the sections above. This section will give an overview and finish with some quotes from the test takers comments.

The information seek task was not as appreciated in the paper mode as in the computer mode but nevertheless received high proportion correct. The matching task was given a higher appraisal in computer mode compared with the paper-based test. The extensive reading task received more or less the same values in paper-based and computer-based delivery. However, it got a higher alpha and the test takers believed they did better on the task compared to the paper-based version (see Table 6.25).

Table 6.25 Cronbach's Alpha, proportion correct, and attitude for PB and CB on each task

Task	Cronbach's Alpha (PB)	Proportion correct (PB)	Attitude (*It was a good test*) (PB)	Attitude (*I think I did well*) (PB)	Cronbach's Alpha (CB)	Proportion correct (CB)	Attitude (*It was a good test*) (CB)	Attitude (*I think I did well*) (CB)
Anchor	.85	.69	-	-	.81	.77	-	-
Matching	.89	.60	3.44	3.23	.84	.64	3.80	3.81
Info. seek	.86	.70	3.37	3.08	.80	.83	3.65	3.62
Gap	.89	.57	3.84	3.43	.84	.64	3.84	3.73
Ext. read	**.83**	**.76**	**3.86**	**3.35**	**.86**	**.77**	**3.86**	**3.61**

To convey deeper insight into the way the test takers perceived the tasks there were some participants who filled in the open comments. Below are a few examples from the test taker feedback for the computer-based tests on testlet A.

> *"I wasn't able to scroll."*
> *"I think it was complicated to start up the test."*
> *"Better than good."*
> *"It was a good one because there were many left when you had 1 left then you couldn't just put it in you had to know or you could guess. There was a glitch that I couldn't scroll to the side then it glitched all the time."*

The next quotations are from testlet A, paper-based version.

> *"i thought it was good but hard."*
> *"It should have been a shorter text."*
> *"It was complicated."*

The next quotations are from testlet B, computer-based version.

> *"Easier to read on paper."*
> *"It's faster to write on a computer."*
> *"I'm more used to paper and pen and it feels safer that way, it's easier to see what you have written and how much you have left."*

"I thought it was pretty fun and interesting assignment."

The next quotations are from testlet B, paper-based version.

"Hard to find the exact fact."

"This text was also very fun to read. Sometimes I thought there were several words that would fit, but that wasn't too big of a problem."

"Well I am just not a big fan of these kind of tests"

6.4 Summary of results

In view of the results displayed above based on a quasi-experimental study some general conclusions can be drawn. The result is valid in the context of English reading comprehension tasks in the Swedish high-stakes national tests and the item formats tested in this study. There were initial differences between the groups of test takers as the groups were not randomized and comparable, corrections were carried out and variables such as grade and anchor scores were included in the analyses. The score reliability was high for both delivery modes and did not differ to any larger extent depending on delivery mode. The item difficulty was not changed between paper-based or computer-based mode and there is score comparability between the two delivery modes. The constructed responses given by the test takers in the computer-based tests were lengthier than those on the paper-based test but they did not achieve higher scores. This will need to be investigated further.

The test takers in testlet B, the three shorter reading comprehension tasks, scored higher on the computer-based test compared to the paper-based test. It was also shown that boys scored higher than girls which is in line with previous research. However, in testlet A, the extensive reading comprehension task, the scores did not differ depending on delivery mode.

The results show different differences between the genders in the two delivery modes. Boys scored higher compared to girls and they gained a higher score, almost 4 points, compared to girls when the test was administered on a computer (testlet B) and a half point on the extensive reading comprehension task (testlet A). However, given that the test takers had the same grade they scored the same on the test.

Test taker feedback shows that the test taker's perception of the task was fairly similar in the extensive reading comprehension task and the gap task but the matching task and the information seek task were more appreciated in the computer-based test compared to the paper-based mode.

The general hypothesis, the null-hypothesis, is that the delivery mode will neither affect the performance of the test takers, nor the reliability of the test score. The results do not give any alarming indications but serve as food for thought. Further research can address questions which have arisen. The following chapter will discuss some of them and possible implications.

7 Discussion and conclusion

The present study was designed to investigate whether two different modes of delivering the same English reading comprehension tests produce the same result and if the scores are comparable. Four groups of students were administered the test either on paper or digitally through a computer. The scores of the groups were then compared with respect to the following research questions.

RQ1. Do the mean score and variance change depending on delivery mode?

Does reliability change depending on delivery mode?

Is item difficulty equal between delivery modes?

RQ2. How will the delivery mode affect the constructive responses given by the test takers?

RQ3. Is the test-taker’s perception of the task affected by the delivery mode?

RQ4. Are there different differences between boys and girls in the two delivery modes?

In this chapter, the results are discussed in the light of previous research. This is followed by some possible implications for high-stakes testing. Limitations of the study are highlighted, this is then followed by some points for recommendations for future research. Finally, concluding remarks are made.

7.1 Answers to the research questions

This section will answer the research questions one at a time and the findings will be discussed in connection to previous research.

- Do the mean score and variance change depending on delivery mode?

The short answer would be no. The first research question looked at whether delivery mode affects the mean score on the tests delivered in two different modes and no substantial proof was found in this study. For testlet A, the extensive reading comprehension task, the mean score was slightly higher for the computer- based test but there was no significant difference (M=11.5 PB; 11.8 CB). The standard deviation was fairly similar (Std. dev. 4.13 PB; 4.48 CB). Nevertheless, causes for potential differences have been suggested by other research. It has for example been suggested that reading on screen is harder and more strenuous than reading on paper (Köpper *et al.* 2016) and that it is more difficult to view the whole text on screen and also fulfil an item (Pommerich 2004). Although my study did not reveal any statistically significant differences between the two modes for the testlet A, more research is needed before any strong conclusions can be drawn.

When analysing data from testlet B. The results show that the mean score for testlet B was slightly higher for the computer-based test (M=24.6 PB; 27.3CB). However, the standard deviation show a wider distribution for the PB-test (11.89 PB; 10.57 CB). The three tasks included in testlet B were not as extensive as the text in testlet A and did not require as much scrolling, nor were the texts included as long as the text in Testlet A. Stepwise regression analysis with the variables “grade” and “mode” showed that a change in delivery mode from paper-based to computer-based would increase the mean score with .318 points. This was not a statistically significant difference. Thus, it could be assumed that delivery mode does not make a significant difference on the mean score in this study. Nevertheless, as mentioned previously, more research is needed into e.g. gender differences and the impact depending on delivery device. In testlet B, three different tasks were included. When examining the mean of each of these tasks, no major differences was noted in the matching task (M=6.7 PB; 7.7 CB) or the gap task (M=8.10 PB; 8.8 CB). However, the information seek task had a higher mean difference (M=9.9 PB; 11.6 CB).

Could a possible explanation to this fact be that this task showed greater differences in layout when going from paper to screen than the other two tasks?

In conclusion, it was found that the mean scores in this study are comparable. In the study made by Karay *et al*. (2015) the conclusion was also that the mean scores did not differ significantly. These findings agree with most studies reviewed in this paper (Porion *et al*. 2016; Singer & Alexander 2017; Brunfaut *et al*. 2018; Endres 2012; Retnawati 2015; Khoshsima *et al*. 2017; Karay *et al*. 2015; Anakwe 2010; Wang *et al*. 2008¸ Barkaoui & Knouzi 2018). However, Maguire *et al*. (2010) found in their study that the scores were significantly higher for the group taking the computer-based test. Jin & Yan (2017) found that students with moderate and high familiarity with computers were scored higher in their writing tests than those less familiar with using a computer. These results were contradicted by the findings in the study by Hosseini *et al.* (2014) who also saw that the impact of the test taker's attitude to using a computer had no significant influence on the score. Further research could investigate the impact of computer familiarity, attitude, and interface differences between different screens on digital devices and that of the paper versus the computer.

- Does the reliability change depending on delivery mode?

A short answer would be no, not in this study. Reliability analysis showed similar Cronbach's alpha values for testlet A. The paper-based test had a Cronbach's alpha value of .77 compared to .76 for the computer-based test. Similarly, no difference in reliability was found for testlet B with Cronbach's alpha of .84 for the paper-based test and .84 for the computer-based test. Neither did a closer look into item reliability show any major differences in reliability coefficients.

A closer look at each of the different tasks included in this study show the following Cronbach's alpha values: anchor task .85 (PB-test) and .81 (CB-test); matching task .89 (PB-test) and .84 (CB-test); info seek task .86 (PB-test) and .80 (CB-test); gap task .89 (PB-test) and .84 (CB-test); extensive reading task .83 (PB-test) and .86 (CB-test). Only one of the tasks, the extensive reading comprehension task, showed a higher the alpha value higher for the computer-based test. These findings are in agreement with the study made by Retnawati (2015), who used classical test theory and found that the two versions had more or less the same reliability coefficients.

- Is item difficulty equal between delivery modes?

Yes, the item difficulty does not change depending on delivery mode. Stepwise regression analysis investigating mode, grade and gender showed that delivery mode does not play a significant role and thus there is comparability between the scores.

However, there were some differences between the four tasks included in the testlets. Mode differences became more apparent in testlet A where the test takers took a more extensive reading comprehension task consisting of an excerpt from a novel for adolescents. The fact that this task required more effort when it comes to reading compared to the three shorter reading comprehension tasks could have an impact on the score. According to Pommerich (2004), reading on screen can be perceived as more difficult than reading on paper since reading on screen includes higher cognitive load and stronger symptoms of eyestrain when longer parts of texts are processed. The fact that reading on screen was perceived to be more difficult was also reported by Köpper *et al*. (2016) and Singer & Alexander (2017) who in their respective studies found that the reader has to scroll to navigate through the text which made reading harder compared to reading on paper.

- How will the delivery mode affect the constructive responses given by the test takers?

When addressing the research question above it became obvious that the delivery mode had an impact on the responses given in this study. An interesting difference in this study was noticed between the two modes on constructed response items requiring the respondents to formulate a more extensive response in writing. On average the responses given on the computer-based test showed lengthier sentences compared to the hand written responses on the paper-based test. This difference did however not result in any difference in score-points nor in measures of proportion correct indicating the item difficulty (the computer-based version had a proportion correct of .77 and the paper-based version .76). Similar differences were also noted by Jin & Yan (2017) who found that participants produced longer sentences on the computer regardless of computer familiarity. Also Barkaoui & Knouzi (2018) found that the test takers on the computer-based test wrote lengthier texts compared to the paper-based mode. They concluded that the better keyboarding skills the test takers had, the lengthier texts they produced. Lengthier responses could thus be the result of the ease of using the keyboard to produce text. Brunfaut *et al*. (2018) report that the test takers in their studies said that it was easier to revise and edit on a computer. The shortcut keyboard commands for copy and paste when writing on a computer might play an important role in this context. It is however also possible to copy text when taking a test on paper but copying texts by hand requires more work than executing shortcut commands on a key board. However, the responses given by the test takers in this study did not appear to be copied to any larger extent. In the light of these findings, further research could give important information. Future research might investigate possible impact on given responses and test scores if there was a limit of number of words possible to enter in a response the test takers of a computer-based test.

- Is the test-taker’s perception of the task affected by the delivery mode?

The answer is twofold, yes and no. The final research question into the perception of the test takers is answered by the analysis of the test taker feedback submitted by the test takers after taking the test. The test takers of the computer-based test scored both the matching task and the info seek higher on the value scale than did the test takers of the paper-based test. There were no differences on the gap task or on the extensive reading task with respect to how the tasks were perceived. However, test takers of the computer-based test believed they achieved better than did the test takers of the paper-based test. Similar results were also observed in several of the studies included in the literature review where the findings show that the participants of their studies perceived that they had a better outcome than they actually had and that they preferred the computer-based test (Mangen *et al.* 2013; Endres 2012; Singer & Alexander 2017; Brunfaut *et al.* 2018; Hosseini 2014).

In the open comments in the test taker feedback, several test takers reported problems with logging in to take the computer-based test. A similar result was found in the study by Retnawati (2015) who saw that less familiarity with computers caused problems logging in, using the mouse and similar technological problems. Time related issues were discussed by Karay *et al.* (2015) who found that test takers responded much faster in the computer-based test. Factors like this could have an impact on the test and need to be investigated further. On the other hand, Bayazit & Aşkar (2012) found that test taker took longer time on a computer-based test compared to a paper-based tests. This is something Alderson (2000) discussed when he argued that time allotment might have to be adjusted when a new delivery mode replaces an old. Pommerich (2004) discussed possible mode effects that might occur with a change in delivery mode. Maybe the test takers could be more motivated when taking a computer-based test compared to one on paper. The endurance might be affected on the extensive reading comprehension task but shorter reading comprehension tasks might build endurance.

- Are there different differences between boys and girls in the two delivery modes?

When considering how boys and girls performed and whether any gender is advantaged by a change in delivery mode, the regression analysis showed that the mean score for boys increased when tests were taken on a computer compared to on paper. This finding is in agreement by Jeong (2012) who found that male test takers achieved higher scores than female test takers when taking computer-based tests. The conclusion Jeong argued for was the impact of higher ICT familiarity within the male group. This fact is supported by Rasmusson (2014a) who found that boys scored higher on digital reading compared to girls and by Rasmusson & Åberg-Bengtsson (2015) who showed that boys had higher visual-spatial skills due to their amount of gaming on the computer. Extra-mural impact was also discussed by Olsson & Sylvén (2015) who found that males showed larger academic vocabulary skills compared to girls. However, the extensive reading task used in this study did not include any hyper-links or skills used in gaming. These results do not agree with the findings from the study by Khoshsima *et al.* (2017) who saw no difference between the scores for boys and girls. This is in contrast with findings form the PISA assessments from 2000 and 2009 where girls in general scored higher compared to boys (SOU 2012:10). Thus, in view of previous research and the results in this study, it is noticeable that gender matters when it comes to computer-based tests. Boys have been found to have higher English language proficiency (Jeong, 2012; Rasmusson 2014a; Rasmusson & Åberg-Bengtsson 2015; Olsson & Sylvén 2015) and it could be noticed that in the testlet with shorter reading comprehension tasks the boys were benefited from the delivery mode. This fact was mentioned as a possible result when the English national tests change delivery mode by the SOU (2016:25). Also Martin & Binkley (2009) discussed the possibility that computer-based reading comprehension tests could decrease the reading literacy gap in favour for boys. A question to investigate is whether longer testing time would benefit girls if it is a question of computer familiarity which gives the boys extra assistance. The missing data could also have contributed to the benefit for boys. In the self-selection of the participating groups and the number of missing there might have been a higher number of boys unfamiliar with computers which could have altered the results in this study. It needs to be investigated if this bias for boys is consistent over time and in other tasks as well.

7.2 Limitations

The validity of the conclusions drawn from the finding in this study have to be combined with the fact the result presented is only representative for this sample and these tasks. Some general conclusions can be stated but the findings are relevant to testing English reading comprehension in the Swedish context and the test item formats used in this study. Due to the size and the representativeness of the sample, this study should mainly be seen as exploratory. One of the most challenging tasks when designing the study was to achieve comparable groups for the analysis since the large-scale pilot tests used in this study were not designed originally for this purpose. The fact that schools participating in the large-scale pilot tests could choose to participate in the computer-based version or the paper-based version may have contributed to a certain self-selection by the schools. This fact could be the answer to the high *p* level of the anchor task for the computer-based test group (.77) compared to the group taking the test on paper (.69).

As mentioned above, the data for this paper came from pilot studies of the test items included in national tests of English. These pilot studies were not designed to address factors that might be influencing test scores, but future pilot tests could benefit from including additional variables such as indicators of computer familiarity, as also suggested by Weigle (2002).

The distribution was higher in the paper-based test for all the tasks except for the extensive reading comprehension task. However, the differences found were not big and these could be explained by the sample groups and the self-selection which occurred when schools could choose to participate and in which mode. Furthermore, this self-selection was augmented by the amount of missing from the computer-based tests due to technical problems and other issues mentioned in the method chapter (ch.5).

Issues concerning the interface could have had an impact on the scores in the computer-based test. This was not investigated further. A future study could investigate whether scores are comparable if test takers have used different devices when taking computer-based tests.

The possible findings in this study were limited to Classical Test Theory. Further analyses using Item Response Theory may have contributed to the result.

7.3 Summary of findings

Thus, these limitations notwithstanding, one of the most important conclusions drawn from this study is that there were no significant differences found in the scores on the two delivery modes, paper-based and computer-based. However, when looking at the open constructed responses the test takers of the computer-based test gave lengthier responses compared to the paper-based test. As digital literacy increases and test takers will become more familiar with using a computer and keyboard instead of paper and pen, a new delivery mode might cause some bias for certain groups of test takers whereas others might find the paper-based delivery mode as more complicated. As computer familiarity increases students will become more comfortable with giving responses on a computer instead of writing with a pencil which is often referred to as more strenuous by the test takers.

Looking at the difference between the means of the total test score of the two modes it was apparent that the group taking the computer-based test had higher English language proficiency compared to the other group. The use of regression analysis made comparisons possible.

It was apparent that boys were favoured by the delivery mode on the computer-based test compared to the paper-based. This might indicate a bias and more research is needed in this area. The mean score on the paper-based test was 25.9 and on the computer-based test 26.3 for boys. This was even more true for the girls when considering that the mean score rose with almost 2 points from paper-based test to computer-based test. The fact that girls had higher grades compared to boys in the summative grading by the teachers as pointed out by Klapp (2015) was apparent when looking at the mean grades between the two genders. All four participating groups in this study had higher mean grades for girls.

The results from this limited study and their implications compared with the finding in the literature review show that there is comparability of scores between the two modes on the tasks included. However, as pointed out by Pommerich (2004), when new digital platforms are used for computer-based testing there is a need for validation studies to establish whether a new construct is tested. New issues which did not occur on paper-based test could arise, e.g. there might be an increase in the time needed to open, complete and hand in a computer-based test. The impact of the test takers digital literacy could thus play an important role on the scores when taking a computer-based test. This was argued by Hughes (2003) when he explained that construct-irrelevant difficulty or easiness could occur with a new delivery mode.

7.4 Implications for practice

No major implications should be drawn from this study as the design is quasi-experimental and the participants were not randomly selected. Similar research into the same tasks but with other participants could come to different finding. Nevertheless, considering these limitations mentioned above and the possible analyses made on the collected data it is possible to draw some implications. The findings show that there is score comparability between the two delivery modes and there appear to be no new construct added to the test. Boys tend to benefit from the new delivery mode on shorter reading comprehension tasks, something which needs to be investigated if it remains the case in among other participating groups. One of the most interesting findings is that test takers produce lengthier responses in constructive response items on the computer-based test. This could have implications on time allotment for the test.

7.5 Points for Future Research

The analyses in this paper were based on classical test theory only. Future research would gain more in-depth knowledge if Item Response Theory was applied as well. Moreover, a deeper analysis is needed of both situational school variables and test taker computer familiarity to differentiate such potential influences from those that may emanate from the delivery mode of the test. Including variables such as questions on computer familiarity and attitudes would give more information to use in future analyses within the NAFS-project.

It is important to investigate possible bias for boys when tests are delivered on a computer.

To gain more knowledge of the impact of the fact that test takers give lengthier response on computer-based test future research might investigate possible impact on given responses and test scores if there was a limit of number of words possible to enter in a response the test takers of a computer-based test.

Another interesting field for research is time related issues. In tests administered via a computer it is possible to gain more information than was possible for paper-based tests. The speed of how fast a response is given can give an indication of the test takers' language proficiency and also indicate if test takers are guessing. Such issues were discussed by Karay *et al.* (2015) who found that test takers responded much faster in the computer-based test.

7.6 Conclusion

The intent of this thesis was to look at the comparability of computer-based test and paper-based test and not to advocate one over the other. The results of this study indicate that there is comparability between the scores when a paper-based test transitions into a computer-based environment and there is high score reliability. Nevertheless, as new formats and new digital platforms and programs are developed these need to be validated to maintain reliability.

There are several implications which can be drawn from these results. Firstly, the results indicate that boys benefit more from computer-based delivery mode when taking shorter reading comprehension tasks and that the benefit is not so significant in the extensive reading comprehension task. Secondly, there might be implications for extensive reading comprehensions tests due to the fact that reading on screen and navigating in a text with scrolling is not as easy as reading on paper. Thirdly, the fact that test takers tend to give lengthier responses on the computer-based constructed items may have some impact on the score and the test taker. And finally, this study did not investigate factors outside the test which could have had an impact on the result and the number of missing in the data.

Reference list

Alderson J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Anakwe, B. (2010). Comparison of Student Performance in Paper-Based Versus Computer-Based Testing. *Journal of Education for Business,*

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bachman, L.F. (2000). Modern language tesing at the turn of the century: Assuring that what we count counts. *Language testing,* 17(1), 1-42.

Barkaoui, K. & Knouzi, I. (2018). (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing*, Vol. 36, pp. 19-31. Available at: https://doi.org/10.1016/j.asw.2018.02.005

Bayazit, A. & Aşkar, P. (2012). *Performance and duration differences between online and paper–pencil tests*. Asia Pacific Educ. Rev. 13: 219. Retrieved the 20171122 from: https://doi-org.ezproxy.ub.gu.se/10.1007/s12564-011-9190-9

Boyd, E. & Taylor, C. (2016). Presenting Validity Evidence: The Case of the *GESE* (p.37-59) In *Contemporary Second Language Assessment.* Vol. 4. Bloomsbury. Edited by Dina Tsagari and Jayanti Banerjee.

Brunfaut, T.; Harding, L. & Batty, A. O. (2018). Going online: The effect of mode of delivery on performance and perceptions on an English L2 writing test suite. *Assessing Writing*, Vol. 36, pp. 3-18. Available at: https://doi.org/10.1016/j.asw.2018.02.003

Buckingham, D. (2008). What Do Young People Need to Know About Digital Media? Colin & Knobel, Michele (Ed.), *Digital literacies: Concepts, Policies and Practices*. New York: Peter Lang Publishing, Inc.

Bugbee, Jr., A.C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education.* Spring96, Vol. 28, Issue 3, p228. 8 p. Retrieved the 20180211 from http://search.ebscohost.com.ezproxy.ub.gu.se/login.aspx?direct=true&db=ehh&AN=9605221096&site=ehost-live

Cazden, C., Cope, B., Fairclough, N., Gee, J., & al, e. The New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, *66*(1), 60. Retrieved *2017-12-04* from https://search-proquest-com.ezproxy.ub.gu.se/docview/212258378?accountid=11162

Chapelle, C. (2007). Technology and Second Language Acquisition. *Annual Review of Applied Linguistics*, 27, 98-114.

Chapelle, C. (2010). Computer-Assisted Teaching and Testing. I Long, M.H. & Doughty, C.J. (Red.), *The Handbook of Language Teaching* (s. 628–644). Chichester: Wiley-Blackwell.

Chapelle, C. & Douglas, D. (2006). Assessing Language through Computer Technology. Cambridge: Cambridge University Press.

Chapelle, C & Voss, E. (2016). 20 Years of Technology and Language Assessment in Language Learning & Technology. *Language Learning & Technology, volume 20, Number 2*, 116-128. Retrieved 20171120 from: http://llt.msu.edu/issues/june2016/chapellevoss.pdf

Chalhoub-Deville, M. (2001). Language Testing and Technology: Past and Future. *Language Learning & Technology*, vol. 5, num. 2 (pp.95-98).

Coleman, J. S. (1988). Social capital in the creation of human capital. American Journal of Sociology, 94, S95–S120.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with new Descriptors.* Retrieved from: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Davey, T. (2011). Practical considerations in computer-based testing. ETS white paper. Princeton, NJ: Educational Testing Service. Retrieved 20171120 from: http://www.ets.org/Media/Research/pdf/CB-2011.pdf.

Douglas, D. (2000) *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Douglas, D. & Hegelheimer, V. (2007) Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27, 115-132.

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder Education.

Endres, H. (2012). A comparability study of computer-based and paper-based Writing Tests. *Research Notes. Cambridge ESOL*, issue 49, pp. 26-32.

Erickson, G. (2004 – revised 2017). *National assessment of foreign languages in Sweden.* Retrieved 20180310 from https://nafs.gu.se/digitalAssets/1671/1671355_national_assessm_of_foreign_lang_in_sweden2017.pdf

Erickson, G. (2009a). *”Att bäras åt” — Om den goda bedömningens flerfaldighet och ömsesidighet*. I U.Tornberg, et al. Språkdidaktiska perspektiv. Om lärande och undervisning i främmande språk (s. 159‒174). Stockholm: Liber.

Erickson, G. (2009b). *Nationella prov i engelska – en studie av bedömarsamstämmighet*. Hämtat 20171123 från http://www.nafs.gu.se/publikationer/

European Council. (2006). *Recommendation of the European Parliament and of the Council of 18 December 2006 on Key Competences for Lifelong Learning.* Official Journal of the European Union. Retrieved *2017-12-04* from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32006H0962

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics.* 5th edition. London: SAGE.

Gee, J. P. (2003). *What Video Games Have to Teach Us about Learning and Literacy.* New York: Palgrave Macmillan.

Gillen, J. (2014). *Digital Literacies.* London: Routledge.

Gipps, C. & Stobart, G. (2009) Chapter 6 *Fairness in Assessment*. Published in *Educational Assessment in the 21st Century. Connecting Theory and Practice.* Editor: Clarie Wyatt-Smith & J. Joy Cumming. Springer. New York.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge Handbooks for LanguageTeachers. (Second Edition). Cambridge: Cambridge University Press.

Internetstiftelsen (2017). *Svenskarna och Internet2017. Undersökning om Svenskarnas Internetvanor.* IIS. Retrieved the 20180315 from https://www.iis.se/docs/Svenskarna_och_internet_2017.pdf

Jeong, H. (2012). A Comparative Study of Scores on Computer-Based Tests and Paper-Based Tests. *Behaviour & Information Technology*, 33:4, pp. 410-422. Available at: https://doi.org/10.1080/0144929X.2012.710647

Jin, Y. & Yan, M. (2017). Computer Literacy and the Construct Validity of a High-Stakes Computer-Based Writing Assessment. *Language Assessment Quarterly*, 14:2, pp. 101-119. Available at: https://doi.org/10.1080/15434303.2016.1261293

Karay, Y., Schauber, S.K., Stosch, C. & Scüttpelz-Brauns, K. (2015). Computer Versus Paper – Does It Make Any Difference in Test Performance? *Teaching and Learning in Medicine*, 27:1, pp. 57-62. Available at: https://doi.org/10.1080/10401334.2014.979175

Khoshsima, H.; Hosseini, M. & Toroujeni, S. M. H. (2017). Cross-Mode Comparability of Computer-based Testing (CB) Versus Paper-Pencil Based Testing (PB): An Investigation of Testing Ad-ministration Mode among Iranian Inter-mediate EFL Learners. *English Language Teaching*, Vol. 10, No. 2.

Klapp, A. (2015). *Bedömning, betyg och lärande.* Lund: Studentlitteratur.

Kline, T.J.B. (2005) *Psychological testing – A Practical Approach to Design and Evaluation*. London: Sage Publishing.

Knobel, M. & Lankshear , C, Editors. (2008). *Digital literacies: Concepts, Policies and Practicies*. New York: Peter Lang Publishing, Inc.

Kozma, R. (2009) Transforming Education: Assessing and Teaching 21st Century Skills. Assessment Call to Action. Published in *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing.* Editors Scheuermann, F. & Björnsson, J. JRC Scientific and Technical Reports. EUR 23679 EN-2009. European Commission.

Kyllonen, P. (2009). New Constructs, Methods, & Directions for Computer-Based Assessment. Published in *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing.* Editors Scheuermann, F. & Björnsson, J. JRC Scientific and Technical Reports. EUR 23679 EN- 2009. European Commission.

Köpper, M., Mayr, S. & Buchner, A. (2016). Reading from computer screen versus reading from paper: does it still make a difference? *Ergonomics,* vol. 59, No. 5, 615-632.

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2017). New Literacies: A Dual-Level Theory of the Changing Nature of Literacy, Instruction, and Assessment. *Journal Of Education*, *197*(2), 1-18.

Lindqvist, A-K. (2012). *Datorbaserad bedömning i språk – forskningsläge, erfarenheter och implikationer*. Masteruppsats i ämnesdidaktik. Göteborgs universitet.

Maguire, K.A., Smith, D.A., Brallier, S.A. & Palm, L.J. (2010). Computer-Based Testing: a Comparison of Computer-Based and Paper-and-Pencil Assessment. *Academy of Educational Leadership Journal*, Vol. 14, No. 4, pp. 117-125.

Mangen, A.; Walgermo, B.R. & Brönnick, K. (2013). Reading linear texts on paper versus computer screen: Effect on reading comprehension. *International Journal of Educational Research*, nr 58, pp. 61-68.

Martin, R. & Binkley, M. (2009) Gender Differences in Cognitive Tests: a Consequence of Gender-dependent Preferncences for Specific Information Presentation Formats? Published in: *The Transistion to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing.* Scheuermann F. & Björnsson, J. (Eds.). European Commission. JRC Scientific and Technical Reports.

Messick, S. A. (1989). Validity. I Linn, R.L. (Red.), *Educational Measurement.* (Third edition), (s. 13‒103). New York: American Council on Education/Macmillan.

Noyes, J.M. & Garland, K.J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, 51:9, pp. 1352-1375. Available at: https://doi.org/10.1080/00140130802170387

OECD (2018). PISA 2018 *Draft Analytical Frameworks, May 2016.* Retrieved 20180615 from: https://www.oecd.org/pisa/data/PISA-2018-draft-frameworks.pdf

Olsson, E. & Sylvén, L.K. (2015). *Extramural English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden.* Apples – Journal of Applied Language Studies. Vol. 9, 2, pp- 77-103.

O'Sullivan, B. (2015). *Technical Report – Aptis Test Development Approach. TR/2015/001*. British Council. English Language Assessment Research Group. Retrieved 20171121from: https://www.britishcouncil.org/sites/default/files/tech_001_barry_osullivan_aptis_test_-_v5_0.pdf

Pommerich M. (2004) *Developing Computerized Versions of Paper-and-Pen Tests: Mode Effects for Passage-Based Testes***.** The Journal of Technology, Learning, and Assessment. Vol. 2, Nr 6, February 2004.

Porion, A., Aparicio, X., Megalakaki, O. & Robert, A. (2016). The impact of paper-based versus computerized presentation of text comprehension and memorization. *Computers in Human Behavior*, 54, pp. 569-576.

Prop. 2017/18:14. *Nationella prov – rättvisa, likvärdiga, digitala.* Retrieved from: https://www.regeringen.se/rattsliga-dokument/proposition/2017/09/prop.-20171814/

Rasmusson, M. (2014a). *Reading Paper – Reading Screen*. Nordic Studies in Education. Vol. 35, pp. 3-19.

Rasmusson, M. (2014b) *Det digitala läsandet. Begrepp, processer och resultat.* Akademisk avhandling i pedagogik. Sundsvall: Mittuniversitetet.

Rasmusson, M. & Åberg-Bengtsson, L. (2015) *Does Performance in Digital Reading Relate to Computer Game Playing? A Study of Factor Structure and Gender Patterns in 15-Year-Olds' Reading Literacy Performance*. Scandinavian Journal of Educational Research, 59:6, 691-709.

Retnawati, H. (2015). The Comparison of Accuracy Scores on the Paper and Pencil Testing vs. Computer-Based Testing. *TOJET, The Turkish Online Journal of Educational Technology,* October, Vol. 14, issue 4.

Ripley, M. (2009) Transformational Computer-based Testing. Published in *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing.* Editors Scheuermann, F. & Björnsson, J. JRC Scientific and Technical Reports. EUR 23679 EN- 2009. European Commission.

Russel, M., Goldberg, A. & O'Connor, K. (2011). *Computer-based Testing and Validity: a look back into the future.* Assessment in Education: Principles, Policy & Practice. Vol. 10, No, 3.

Singer, L. & Alexander, P. A. (2017). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. *The Journal of Experimental Education*, 85:1, pp. 155-172. Available at: https://doi.org/10.1080/00220973.2016.1143794

Skolverket. (2017a). *Kommentarmaterial till kursplanen i engelska.* Stockholm. Edita. Retrieved from https://www.skolverket.se/sitevision/proxy/publikationer/svid12_5dfee44715d35a5cdfa2899/55935574/wtpub/ws/skolbok/wpubext/trycksak/Blob/pdf3858.pdf?k=3858

Skolverket. (2017b) *Skolverkets sysmtemramverk för nationella prov.* Retrieved from https://www.skolverket.se/sitevision/proxy/publikationer/svid12_5dfee44715d35a5cdfa2899/55935574/wtpub/ws/skolbok/wpubext/trycksak/Blob/pdf3890.pdf?k=3890

SOU 2012:10. *Läsarens marknad, marknadens läsare – en forskningsantologi.* Stockholm: Fritzers Offentliga Publikationer.

SOU 2016:25. *Likvärdigt, rättsäkert och effektivt – ett nytt nationellt system för kunskapsbedömning.* Stockholm: Wolters Kluwers Sverige AB.

Sylvén, L.K. & Sundqvist, P. (2012). *Gaming as extramural English L2 learning an L2 proficiency among young learners*. European Association for Computer Assisted Language Learning. 24(3), pp. 302-321.

Takala, Erickson, Figuera & Gustafsson. (2016). Future Prospects and Challenges in Language Assessments. In *Contemporary Second Language Assessment*. Vol. 4. Bloomsbury 2016. Edited by Dina Tsagari and Jayanti Banerjee.

Sawaki, Y. (2001) Comparability of Conventional and Computerized Tests of Reading in a Second Language. *Language Learning & Technology,* May, Vol. 5, Num. 2 (pp. 38-59).

Singer, L.M. & Alexander, P.A. (2017). Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration. In *The Journal of Experimental Education.* Vol. 85.No.1, 155-177.

UNESCO (2008a) *Understanding Information Literacy: A Primer.* Retrieved *2017-12-04* from: http://www.uis.unesco.org/Communication/Documents/157020E.pdf d.

Vetenskapsrådet. (2017). *God forskningssed.* Vetenskapsrådet. Retrieved the 20170222 from file:///C:/Users/xasplt/Downloads/God-forskningssed-2017%20(1).pdf

Vuorikari, R., Punie, Y., Carretero Gomez S., Van den Brande, G. (2016). *DigComp 2.0: The Digital Competence Framework for Citizens*. Update Phase 1: The Conceptual Reference Model. Luxembourg Publication Office of the European Union. EUR 27948 EN. doi:10.2791/11517

Wang, S., Jiao, H., Young, M.J., Brooks, T., Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K-12 Reading Assessments. *Educational and Psychological Measurement*, Vol. 68, No. 1, pp. 5-24.

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press. Series editors J. Charles Alderson & Lyle F. Bachman 2002

Weir, C. J. (2005). *Language Testing and Validation. An evidence-based Approach.* New York: Palgrave Macmillan.

Appendix 1: Literature used in the Literature Review

Author	Title	Year	Name of Journal	Method/ Type of Publication	Keywords
Brunfaut, T.; Harding, L. & Batty, A. O.	Going online: The effect of mode of delivery on performance and perceptions on an English L2 writing test suite	2018	Assessing Writing	Case study	Paper-based testing of writing; Computer-based testing of writing; Online testing of writing, Mode of delivery; Perceptions; second language writing assessment
Mangen, A.; Walgermo, B.R. & Brönnick, K.	Reading linear texts on paper versus computer screen: Effect on reading comprehension	2013	International Journal of Educational Research	Case study	Reading comprehensio n; Screen reading; Print reading; Computers in education
Endres, H.	A comparability study of computer-based and paper-based Writing Tests	2012	Research Notes. Cambridge ESOL, issue 49.	Qualitative study Case study	Computer-based test, writing tests, Paper-based test, comparability

Khoshsima, H.; Hosseini, M. & Toroujeni, S. M. H.	Cross-Mode Comparability of Computer-based Testing (CB) Versus Paper-Pencil Based Testing (PB): An Investigation of Testing Administration Mode among Iranian Intermediate EFL Learners	2017	English Language Teaching	Case study	Computer-based testing; paper-based testing, gender difference; test preference
Singer, L. & Alexander, P. A.	Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration	2017	The Journal of Experimental Education	Case study	Calibration; computers in education; digital reading; print reading; reading comprehension
Jin, Y. & Yan, M.	Computer Literacy and the Construct Validity of a High-Stakes Computer-Based Writing Assessment	2017	Language Assessment Quarterly	Case study	

Wang, S., Jiao, H., Young, M.J., Brooks, T., Olson, J.	Comparability of Computer-Based and Paper-and-Pencil Testing in K-12 Reading Assessments	2008	Educational and Psychological Measurement	Meta-analysis	Meta-analysis; computer-based testing; comparability of educational test modes; K-12 reading tests
Noyes, J.M. & Garland, K.J.	Computer- vs. paper-based tasks: Are they equivalent?	2008	Ergonomics	Literature review	Computer vs. oaoer; NASA.TLX workload measure; online assessment; performance indices
Anakwe, B.	Comparison of Student Performance in Paper-Based Versus Computer-Based Testing	2010	Journal of Education for Business	Case study	Assessments, in-class; online; tests

Porion, A., Aparicio, X., Megalakaki, O. & Robert, A.	The impact of paper-based versus computerized presentation of text comprehension and memorization	2016	Computers in Human Behavior	Experimental study	Reading; assessment; paper; computer; comprehension; memory
Karay, Y., Schauber, S.K., Stosch, C. & Scüttpelz-Brauns, K.	Computer Versus Paper – Does It Make Any Difference in Test Performance?	2015	Teaching and Learning in Medicine	Case study	Difference in test performance; computer-based test versus paper-based test, formative Progress Test
Retnawati, H.	The Comparison of Accuracy Scores on the Paper and Pencil Testing vs. Computer-Based Testing	2015	The Turkish Online Journal of Educational Technology	Case study	Accuracy; reliability; value of information function (VIF); paper and pencil test (PB); computer-based test (CB).

Jeong, H.	A Comparative Study of Scores on Computer-Based Tests and Paper-Based Tests	2012	Behaviour & Information Technology	Case study	Computer-based test; paper-based test; gender; subject
Maguire, K.A., Smith, D.A., Brallier, S.A. & Palm, L.J.	Computer-Based Testing: a Comparison of Computer-Based and Paper-and-Pencil Assessment	2010	Academy of Educational Leadership Journal	Case study	
Barkaoui, K,	The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores	2018	Assessing Writing	Exploratory study	Writing mode; keyboard skills; second language proficiency; essay linguistic characteristic

Appendix 2 *Test Taker Feedback*

Questionnaire

	Yes, absolutely				**No, absolutely not**
1. The task X was a good	-----------	-----------	-----------	----------	----------
2. It was difficult	-----------	-----------	-----------	----------	----------
3. It was interesting to read	-----------	-----------	-----------	----------	----------
4. There were many words that I didn't understand	-----------	-----------	-----------	----------	----------
5. I think I did well on this part of the test	-----------	-----------	-----------	----------	----------

Comments about X (you can write in English or Swedish)

____

____

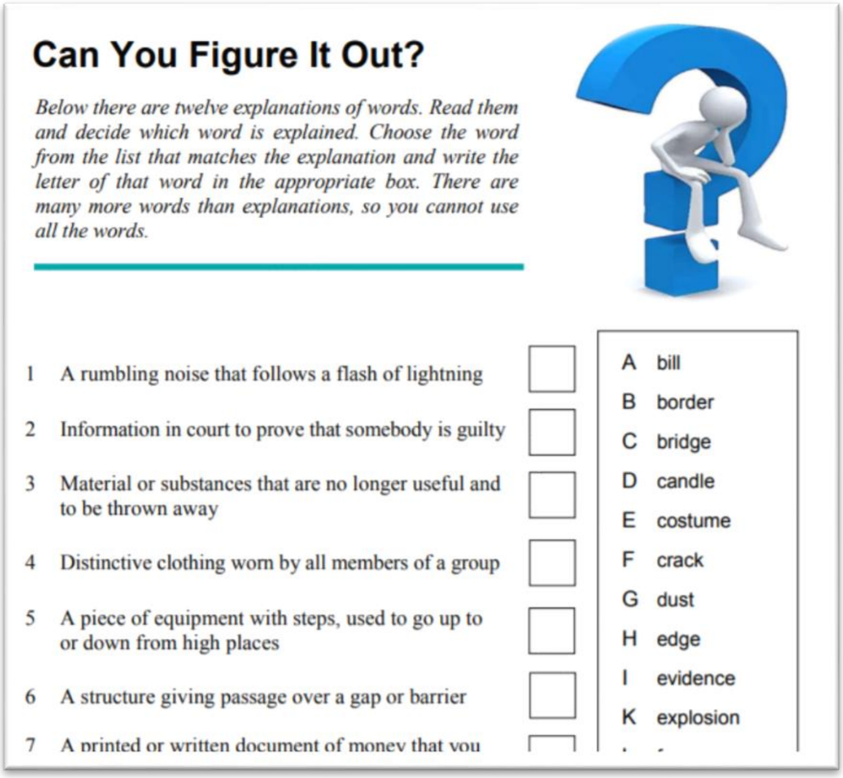
____

____

____

____

Appendix 3 Examples of tasks



Can You Figure It Out?

Below there are twelve explanations of words. Read them and decide which word is explained. Choose the word from the list that matches the explanation and write the letter of that word in the appropriate box. There are many more words than explanations, so you cannot use all the words.

1 A rumbling noise that follows a flash of lightning □

2 Information in court to prove that somebody is guilty □

3 Material or substances that are no longer useful and to be thrown away □

4 Distinctive clothing worn by all members of a group □

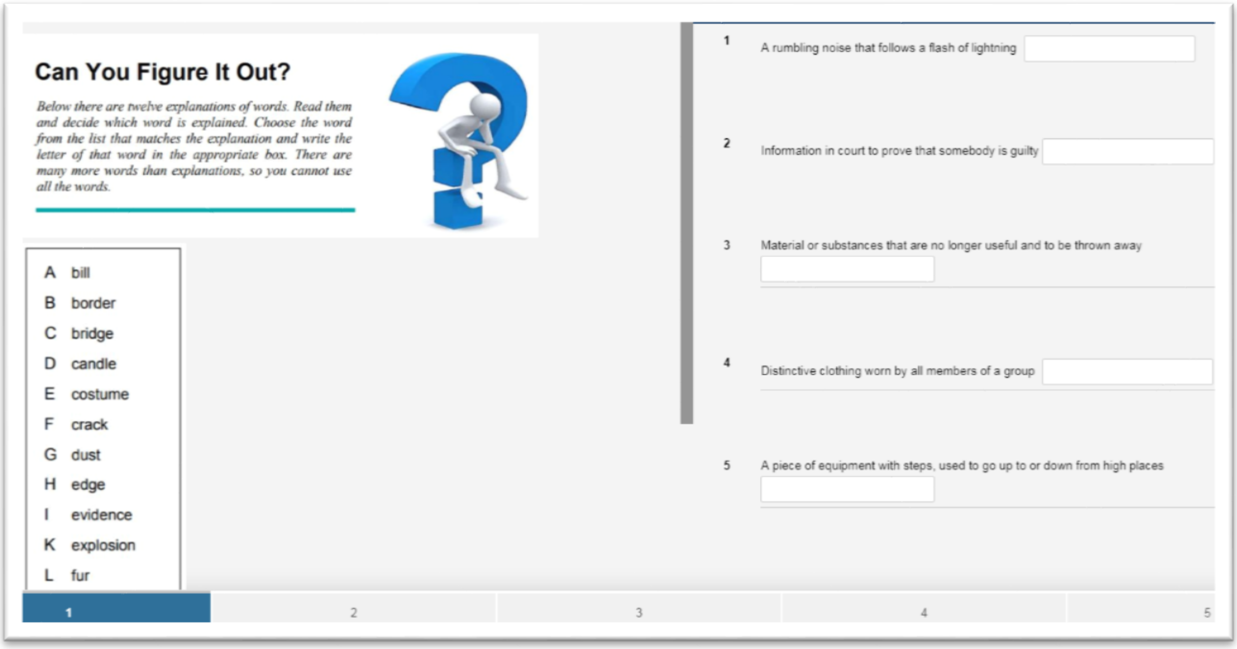
5 A piece of equipment with steps, used to go up to or down from high places □

6 A structure giving passage over a gap or barrier □

7 A printed or written document of money that you □

A bill
B border
C bridge
D candle
E costume
F crack
G dust
H edge
I evidence
K explosion

Fig. 5.1 Matching task, paper delivery.



Can You Figure It Out?

Below there are twelve explanations of words. Read them and decide which word is explained. Choose the word from the list that matches the explanation and write the letter of that word in the appropriate box. There are many more words than explanations, so you cannot use all the words.

A bill
B border
C bridge
D candle
E costume
F crack
G dust
H edge
I evidence
K explosion
L fur

1 A rumbling noise that follows a flash of lightning

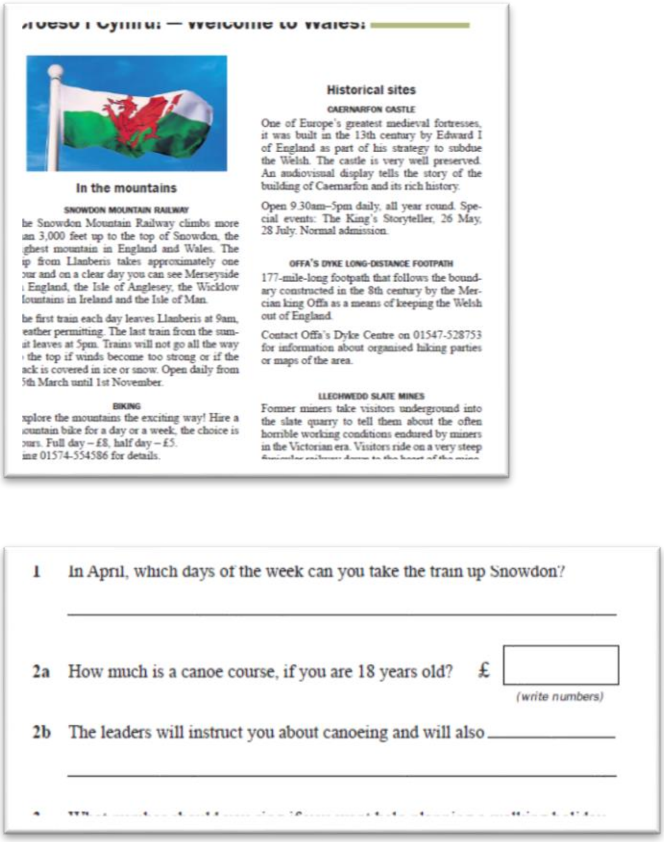
2 Information in court to prove that somebody is guilty

3 Material or substances that are no longer useful and to be thrown away

4 Distinctive clothing worn by all members of a group

5 A piece of equipment with steps, used to go up to or down from high places

Fig 5.2 Matching task, computer delivery.



Croeso i Cymru! — Welcome to Wales!

In the mountains

SNOWDON MOUNTAIN RAILWAY

he Snowdon Mountain Railway climbs more an 3,000 feet up to the top of Snowdon, the ghest mountain in England and Wales. The ip from Llanberis takes approximately one our and on a clear day you can see Merseyside England, the Isle of Anglesey, the Wicklow lountains in Ireland and the Isle of Man.

he first train each day leaves Llanberis at 9am, eather permitting. The last train from the sumit leaves at 5pm. Trains will not go all the way the top if winds become too strong or if the ack is covered in ice or snow. Open daily from 5th March until 1st November.

BIKING

xplore the mountains the exciting way! Hire a ountain bike for a day or a week, the choice is ours. Full day – £8, half day – £5. ing 01574-554586 for details.

Historical sites

CAERNARFON CASTLE

One of Europe's greatest medieval fortresses, it was built in the 13th century by Edward I of England as part of his strategy to subdue the Welsh. The castle is very well preserved. An audiovisual display tells the story of the building of Caernarfon and its rich history.

Open 9.30am–5pm daily, all year round. Special events: The King's Storyteller, 26 May, 28 July. Normal admission.

OFFA'S DYKE LONG-DISTANCE FOOTPATH

177-mile-long footpath that follows the boundary constructed in the 8th century by the Mercian king Offa as a means of keeping the Welsh out of England.

Contact Offa's Dyke Centre on 01547-528753 for information about organised hiking parties or maps of the area.

LLECHWEDD SLATE MINES

Former miners take visitors underground into the slate quarry to tell them about the often horrible working conditions endured by miners in the Victorian era. Visitors ride on a very steep funicular railway down to the heart of the mine

1 In April, which days of the week can you take the train up Snowdon?

____

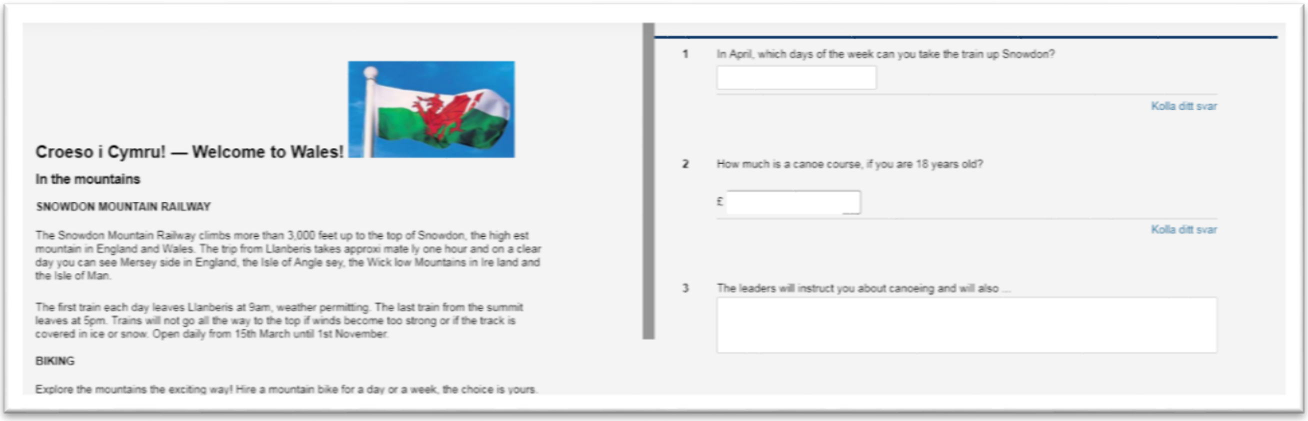
2a How much is a canoe course, if you are 18 years old? £ □

(write numbers)

2b The leaders will instruct you about canoeing and will also ____________

____

Fig 5.3 Information seek task, paper delivery.



Croeso i Cymru! — Welcome to Wales!

In the mountains

SNOWDON MOUNTAIN RAILWAY

The Snowdon Mountain Railway climbs more than 3,000 feet up to the top of Snowdon, the high est mountain in England and Wales. The trip from Llanberis takes approxi mate ly one hour and on a clear day you can see Mersey side in England, the Isle of Angle sey, the Wick low Mountains in Ire land and the Isle of Man.

The first train each day leaves Llanberis at 9am, weather permitting. The last train from the summit leaves at 5pm. Trains will not go all the way to the top if winds become too strong or if the track is covered in ice or snow. Open daily from 15th March until 1st November.

BIKING

Explore the mountains the exciting way! Hire a mountain bike for a day or a week, the choice is yours.

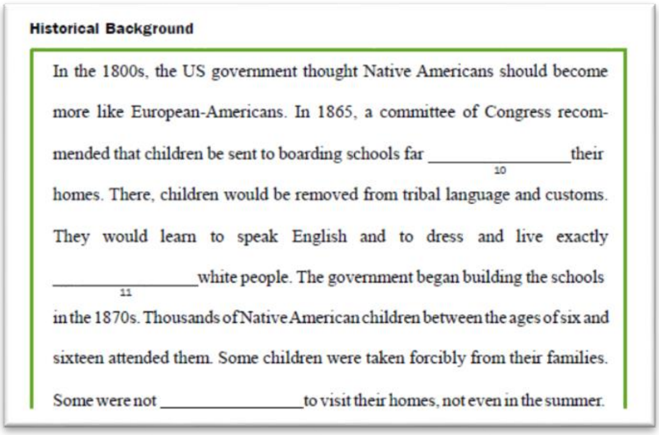
1 In April, which days of the week can you take the train up Snowdon?

2 How much is a canoe course, if you are 18 years old?

£

3 The leaders will instruct you about canoeing and will also ...

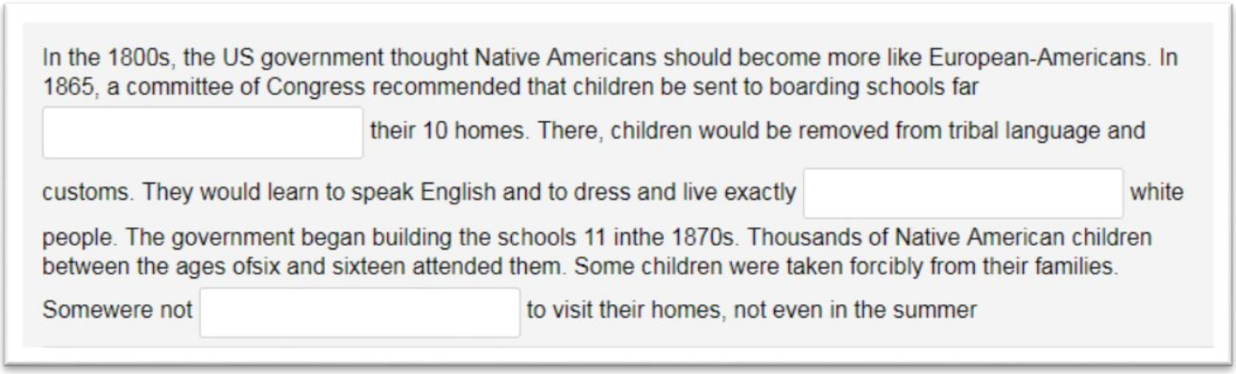
Fig 5.4 Information seek task, computer delivery.



Historical Background

In the 1800s, the US government thought Native Americans should become more like European-Americans. In 1865, a committee of Congress recommended that children be sent to boarding schools far _______________ 10 their homes. There, children would be removed from tribal language and customs. They would learn to speak English and to dress and live exactly _______________ 11 white people. The government began building the schools in the 1870s. Thousands of Native American children between the ages of six and sixteen attended them. Some children were taken forcibly from their families. Some were not _______________ to visit their homes, not even in the summer.

Fig 5.5 Gap task, paper delivery.

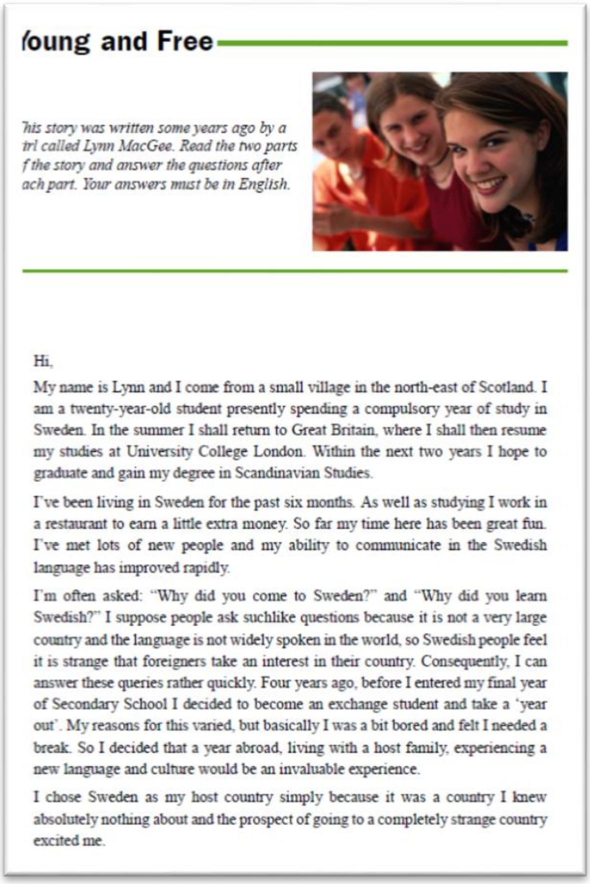


In the 1800s, the US government thought Native Americans should become more like European-Americans. In 1865, a committee of Congress recommended that children be sent to boarding schools far

______________ their 10 homes. There, children would be removed from tribal language and customs. They would learn to speak English and to dress and live exactly ______________ white people. The government began building the schools 11 inthe 1870s. Thousands of Native American children between the ages ofsix and sixteen attended them. Some children were taken forcibly from their families.

Somewere not ______________ to visit their homes, not even in the summer

Fig 5.6 Gap task, computer delivery.



oung and Free

his story was written some years ago by a irl called Lynn MacGee. Read the two parts f the story and answer the questions after ach part. Your answers must be in English.

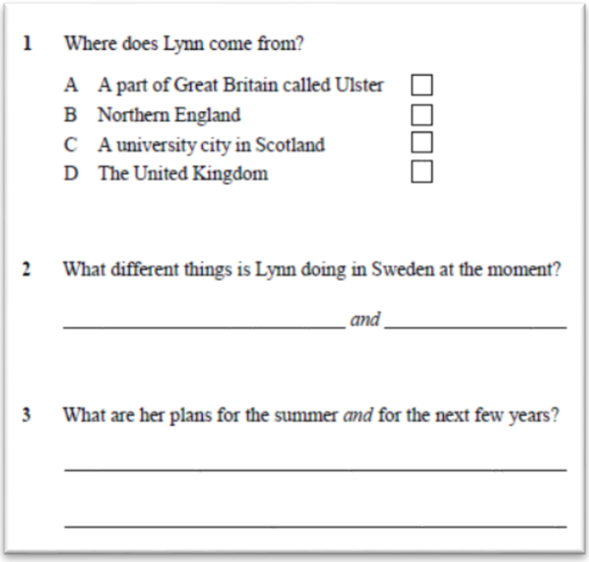
Hi,

My name is Lynn and I come from a small village in the north-east of Scotland. I am a twenty-year-old student presently spending a compulsory year of study in Sweden. In the summer I shall return to Great Britain, where I shall then resume my studies at University College London. Within the next two years I hope to graduate and gain my degree in Scandinavian Studies.

I've been living in Sweden for the past six months. As well as studying I work in a restaurant to earn a little extra money. So far my time here has been great fun. I've met lots of new people and my ability to communicate in the Swedish language has improved rapidly.

I'm often asked: "Why did you come to Sweden?" and "Why did you learn Swedish?" I suppose people ask suchlike questions because it is not a very large country and the language is not widely spoken in the world, so Swedish people feel it is strange that foreigners take an interest in their country. Consequently, I can answer these queries rather quickly. Four years ago, before I entered my final year of Secondary School I decided to become an exchange student and take a 'year out'. My reasons for this varied, but basically I was a bit bored and felt I needed a break. So I decided that a year abroad, living with a host family, experiencing a new language and culture would be an invaluable experience.

I chose Sweden as my host country simply because it was a country I knew absolutely nothing about and the prospect of going to a completely strange country excited me.



1 Where does Lynn come from?

A A part of Great Britain called Ulster □
B Northern England □
C A university city in Scotland □
D The United Kingdom □

2 What different things is Lynn doing in Sweden at the moment?

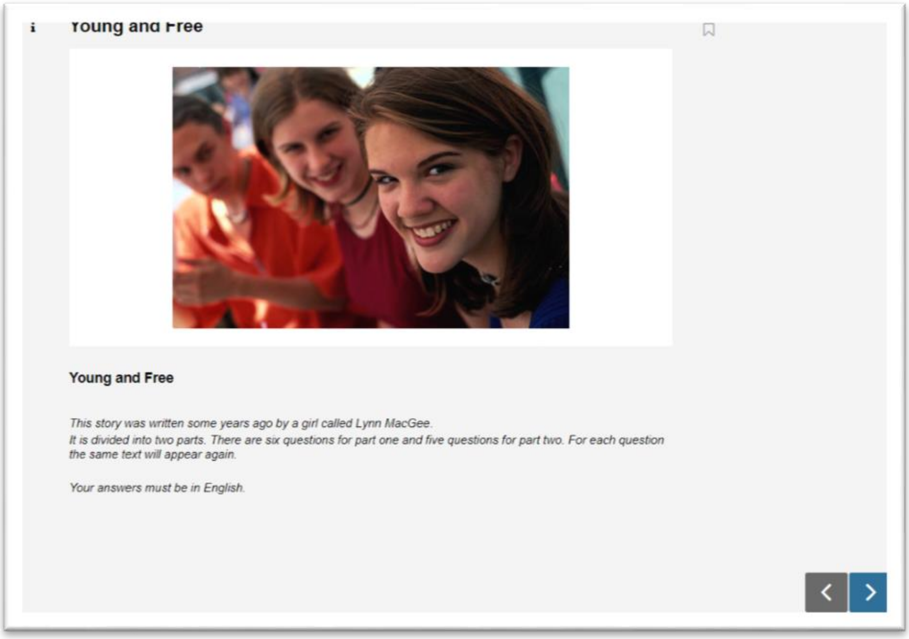
______________________________ *and* ___________________

3 What are her plans for the summer *and* for the next few years?

____

____

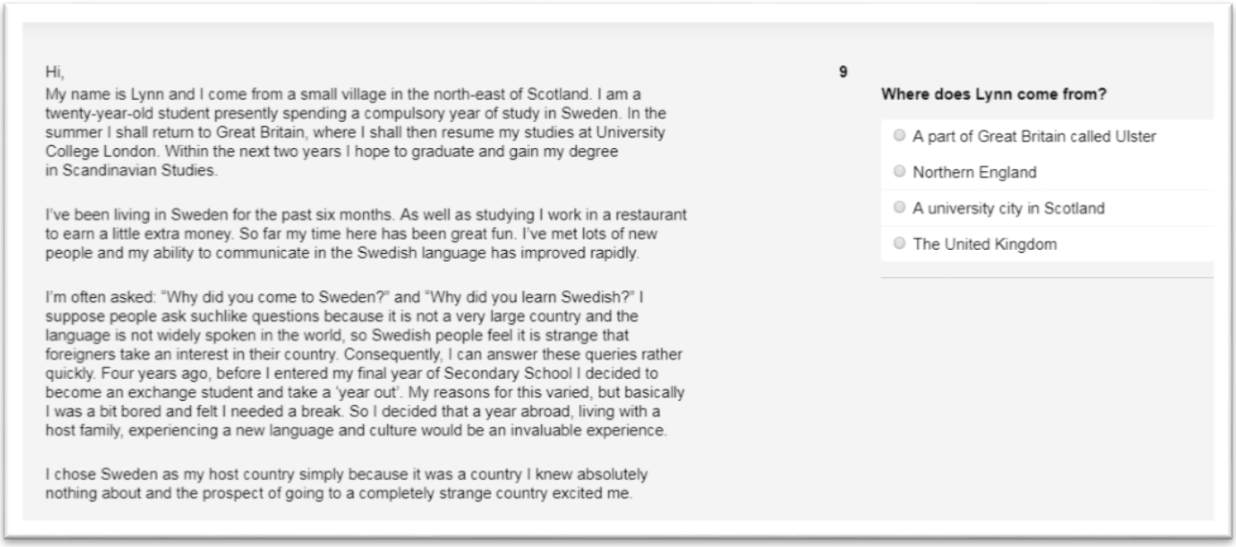
Fig 5.7 Extensvive reading comprehension task, paper delivery.



Young and Free

This story was written some years ago by a girl called Lynn MacGee.
It is divided into two parts. There are six questions for part one and five questions for part two. For each question the same text will appear again.

Your answers must be in English.



Hi,
My name is Lynn and I come from a small village in the north-east of Scotland. I am a twenty-year-old student presently spending a compulsory year of study in Sweden. In the summer I shall return to Great Britain, where I shall then resume my studies at University College London. Within the next two years I hope to graduate and gain my degree in Scandinavian Studies.

I've been living in Sweden for the past six months. As well as studying I work in a restaurant to earn a little extra money. So far my time here has been great fun. I've met lots of new people and my ability to communicate in the Swedish language has improved rapidly.

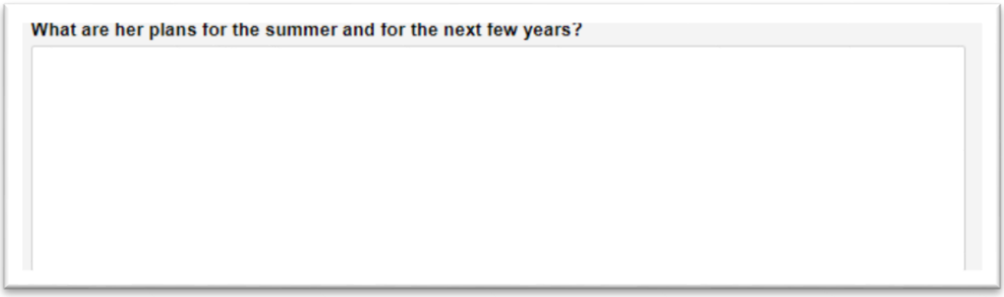
I'm often asked: "Why did you come to Sweden?" and "Why did you learn Swedish?" I suppose people ask suchlike questions because it is not a very large country and the language is not widely spoken in the world, so Swedish people feel it is strange that foreigners take an interest in their country. Consequently, I can answer these queries rather quickly. Four years ago, before I entered my final year of Secondary School I decided to become an exchange student and take a 'year out'. My reasons for this varied, but basically I was a bit bored and felt I needed a break. So I decided that a year abroad, living with a host family, experiencing a new language and culture would be an invaluable experience.

I chose Sweden as my host country simply because it was a country I knew absolutely nothing about and the prospect of going to a completely strange country excited me.

9

Where does Lynn come from?

○ A part of Great Britain called Ulster
○ Northern England
○ A university city in Scotland
○ The United Kingdom



What are her plans for the summer and for the next few years?

Fig 5.8 Extensive reading comprehension task, computer delivery.