Univerzita Karlova

Filozofická fakulta

Ústav anglického jazyka a didaktiky

Didaktika konkrétního jazyka

Linda Nepivodová

Disertační práce

In Their Own Words and By the Test Results: A mixed methods study of the comparison of two modes of test administration

Vlastními slovy studentů a podle výsledků testů: Smíšený výzkum porovnávající dva způsoby administrace testů

Vedoucí práce: Doc. PhDr. Lucie Betáková, MA, Ph.D.

2017

Prohlašuji, že jsem disertační práci napsala samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Brně, dne 28.3.2017

……..
Jméno a příjmení

Poděkování:

Poděkování patří mé školitelce Doc. PhDr. Lucii Betákové, MA, Ph.D. za její vedení a podporu.
Děkuji také Katedře anglistiky a amerikanistiky FF MU za vytvoření prostředí a podmínek, které mně umožnily realizovat výzkum, a Mgr. Jakubovi Chromcovi za statistické konzultace.

Abstract
The thesis compares first year university students' compulsory English achievement tests written in two modes of administration, i.e. pencil and paper and computer-based tests (PPT and CBT), over three successive years in an English department in the Czech Republic (2014-2016). The analysis of the three stages of the project, Pilot, Study 1 and Study 2, forms the focus of the study. The central research question investigated is whether the usage of computer-based language testing is justified in such a context.
The dissertation first maps the theoretical discourse and comments on selected fundamental aspects of language testing practices, significant for the comparison itself. A convergent parallel mixed methods design is employed, where quantitative test data is used to measure the equivalency of the two modes and qualitative self-report feedback form data explores the main advantages and disadvantages of these, as viewed by the students examined.
The results demonstrate the following conclusions: The differences in the scores gained from the CBT and PPT modes in all three stages of the research are statistically insignificant, though overall the students perform slightly better in the PPT modes. No striking differences were found concerning gender either, though the prediction put male students' scores in the CBT modes above the female students' ones. Significant differences as regards the question type were then observed. In both modes students perform better with multiple choice questions, while with short answer items students are more successful in the PPT mode. In the students' feedback forms, Manipulation appears to be the most prominent in all the three stages among the advantages of the PPT mode, covering the possibility of making notes. Other factors changed over the time given; Orientation, for example, which in Pilot and Study 1 was among the most frequently cited advantages of the PPT, ceased to be considered as advantageous in Study 2. The main advantage of the CBT mode as identified by the students over the three years is indisputably the immediacy of results. While in Study 1 the preferences concerning the two modes are almost equally divided, in Study 2, the CBT is preferred to the PPT, however, the category of 'No preference' almost doubled. The statistical model is employed to ascertain whether there are statistically significant differences between the mean subtest scores in connection to student preferences, however, none were found. As concluded, the researcher believes that sufficient evidence has been provided to consider the usage of computer-based achievement tests in the context given to be justified.

Key words: English achievement test, pencil and paper-based test (PPT), computer-based test (CBT), modes of administration, comparison, convergent parallel mixed methods design, quantitative and qualitative data, equivalency, self-report feedback, gender, question type, manipulation, orientation, immediacy of results, preference.

Abstrakt
Tématem dizertace je tříletý výzkumný projekt (2014-2016), zabývající se porovnáním povinných testů pokroku z angličtiny, které jsou zadávány univerzitním studentům prvního ročníku katedry anglistiky v České republice ve dvou podobách, a to v tradiční písemné formě a formě počítačové. Analýza tří fází projektu, Pilot, Study 1 a Study 2, tvoří ústřední část práce. Hlavní otázka výzkumu spočívá v tom, zda je současná praxe používání počítačové formy testu v tomto kontextu opodstatněná.
Dizertace nejprve shrnuje teoretickou debatu a komentuje vybrané podstatné aspekty, důležité pro vlastní srovnání. Zde je aplikována konvergentní paralelní smíšená metoda, kde jsou kvantitativní údaje použity k poměření ekvivalence mezi dvěma způsoby zadávání testu, zatímco kvalitativní údaje reprezentované vlastní zpětnou vazbou jednotlivých studentů osvětlují výhody a nevýhody obou forem testu z jejich úhlu pohledu.
Výsledná interpretace je následující: Výsledky u tradičních písemných i počítačových testů nevykazují ani v jednom stádiu výzkumu žádný statisticky významný rozdíl, ačkoli u písemného testu jsou počty bodů celkově mírně vyšší. Genderové rozdíly nebyly zjištěny, i když se předpokládalo, že studenti mužského pohlaví budou mít lepší skóre u počítačových testů. Významné rozdíly byly pak zjištěny u typu testových otázek. U obou forem testu mají studenti lepší výsledky u volených odpovědí, zatímco u krátkých odpovědí získávají studenti vyšší bodové ohodnocení u písemné formy. U zpětné vazby studentů se aspekt Manipulace významně objevuje mezi nejčastějšími výhodami tradičního písemného testu, který umožňuje dělat si poznámky. Jiné faktory se během výzkumu mění; například Orientaci, která byla uváděna jako jedna z nejčastějších výhod písemné formy u fází Pilot a Study 1, volí jako výhodu u Study 2 jen menšina studentů. Hlavní výhodou počítačové formy testu je pak podle studentů tříletého výzkumu možnost vidět okamžitý výsledek. Zatímco u Study 1 jsou preference ve vztahu k zadané formě testu téměř vyrovnány, u Study 2 preference k počítačové formě již převažují, avšak na rozdíl od Study 1, se kategorie 'Žádná preference' téměř zdvojnásobila. Aplikovaný statistický model zjišťuje, zda existují významné rozdíly mezi výsledkem testu ve vztahu k preferencím studenta. Ty však nebyly nalezeny. V závěru vyjadřuje autorka výzkumu přesvědčení, že bylo shromážděno dostatečné množství důkazů, které v popsaném kontextu plně opravňují používat u testů pokroku počítačovou formu.

Klíčová slova: test pokroku z angličtiny, písemná forma testu, počítačová forma testu, způsoby zadávání, porovnání, konvergentní paralelní smíšená metoda, kvantitativní a kvalitativní data, ekvivalence, vlastní zpětná vazba, gender, typ otázky, manipulace, orientace, okamžitý výsledek, preference.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

LIST OF ABBREVIATIONS

ARG	Assessment Reform Group
CALT	Computer Assisted Language Testing/ Test
CAT	Computer Adaptive Testing/ Test
CB	Computer-based
CBT	Computer-based Test
CEFR	Common European Framework of Reference
CRT	Criterion-referenced Test
EAP	English for Academic Purposes
ESP	English for Specific Purposes
ETS	Educational Testing Service
GRE	Graduate Record Examination
HTML	Hypertext Mark-up Language
IRT	Item Response Theory
M	Mean
MCQ	Multiple Choice Question
N	Number
NLP	Natural Language Processing
NRT	Norm-referenced Test
PISA	Programme for International Student Assessment
PP	Pencil and Paper-based
PPT	Pencil and Paper-based Test
RQ	Research Question
SA	Short Answer
SAT	Scholastic Aptitude Test
SD	Standard Deviation
TEA	Technology Enhanced Assessment
TSP	Test studijních předpokladů (Study Aptitude Test)
UCLES	University of Cambridge Local Examinations Syndicate
WBT	Web-based test

1 INTRODUCTION

'*Teaching and testing are two inseparable parts of the educational process.*'

(Farhady, 2006, p.3)

While teaching is given a fair amount of attention and teachers are well informed and prepared for the what, how, when and who to teach, testing in general is too often neglected and given very little, if any, attention. The aim of the present dissertation is to alter this paradigm and give testing the attention it undoubtedly deserves.

With rapid and vast advances in technology, computer-based language testing has grown in popularity and is being implemented across various contexts in the Czech Republic. However, it remains a heavily under-researched area, at least in the Czech context. The author of this dissertation has examined various aspects of computer-based language testing – at that time in its infancy - in her Master Thesis (2006), in which she carried out an empirical study revealing a number of interesting points, yet at the same time raising multiple questions related to the comparison of computer-based and pencil and paper-based tests.

A strong personal motivation in exploring the area under scrutiny further needs to be emphasized, since the author hopes that the present research will shed light on the advantages and disadvantages of computer-based language testing as viewed by students, lead to improvement in the testing situation at her workplace as well as offer recommendations to institutions who are considering the implementation of computer-based tests.

The context of the dissertation is the tertiary level, i.e. university students, more specifically first year students of English at the Department of English and American studies at the Faculty of Arts, Masaryk University in the Czech Republic. Compulsory achievement tests written in both modes of test administration are going to be studied and analysed in 3 successive years.

The central research question investigated throughout is whether the usage of computer-based language testing is justified in such a context.

1.1 Purpose statement

This mixed methods research addresses potential differences regarding two modes of test administration (pencil and paper-based and computer-based respectively). A convergent parallel mixed methods design will be used. This type of research design focuses on both quantitative and qualitative data analysis with both types of data collected in parallel. Such data is analysed separately and then merged in the overall interpretation of the research problem.

In this study, quantitative test data will be used to measure the equivalency of the two modes of test administration and qualitative self-report feedback form data will explore the main advantages and disadvantages of the two modes as viewed by students.

The reason for collecting both quantitative and qualitative data is to corroborate results in order to bring greater insights into the lesser-investigated area in the Czech context. It is believed that the integration of the two types of data will validate results more efficiently than if obtained by either quantitative or qualitative data separately.

1.2 Structure of the dissertation

The present dissertation consists of two wider sections, the former being based on the discourse concerning selected theoretical issues and the latter comprising the author's research project, its analysis and comment on the criteria examined.

In the second chapter, **Fundamental Concepts in Language Testing**, some of the main considerations in language testing are discussed in order to introduce the theoretical background to the study. *A Brief Historical Context* of language testing practices is outlined and some significant representatives in the area explored in its different stages of development are stated. Specific eras with their categories, divisions, terms, methodologist approaches, techniques and trends are presented, reflecting the changes in testing practices over time.

Significant aspects of validity, reliability, practicality and washback within testing practices are dealt with in the following subchapter, *Test Qualities*. Here, different validities, for example, content, face, response and construct and scoring validities are detailed. The tension between validity and reliability is then highlighted by various scholars' arguments. There are a number of factors that are associated with the practicality of a test, which is heavily influenced by human and material resources, for example. Here, machine scoring is also mentioned, which is significantly relevant for the present research project. Washback in relation to Practicality and Feedback is then contrasted and its positive and negative effects described and discussed.

The following subchapter, *Test Types Dichotomies*, first covers the distinction between Criterion-referenced and Norm-referenced tests and their different functions, which is highly relevant for the present research project. Second, Discrete-point versus Integrative tests are examined, and the third dichotomy includes Summative versus Formative assessments, which concern the product and process of learning. A number of scholars' opinions are manifested here in connection with the above stated dichotomies, polarizing thus both the significant points and arguments. Competence versus Performance is then presented as the last dichotomy, with some scholars accentuating the overall unifying rather than dividing components of the issue.

Computer Assisted Language Testing (CALT) forms the basis of the third chapter. Here, the author draws on her MA dissertation (2006), which shows that the CALT benefits outnumber the drawbacks, demonstrating clearly in view of present day technologies that computers are faster and more efficient in most areas. Various merits and demerits, claimed by different authors, are included in the subchapter entitled *CALT Advantages and Disadvantages*, and the list of these proposed by H. Douglas Brown in 2010 is stated and commented on. Other scholars, such as Chapelle and Douglas (2006), identify within CALT a number of potential validity threats and suggest measures to diminish them.

The CALT chapter is further subdivided into *Digital Literacy*, *Towards a CALT framework*, *Innovative Item Types* and *Comparability Studies*, which provide

detailed discussion based on a number of scholars' arguments and supporting evidence related to their standpoints, in particular when comparing the two modes, pencil and paper-based tests (PPT) and computer-based tests (CBT).

The empirical part begins with **Research Methodology** (Chapter 4), in which the multi-phased mixed methods research design adopted is explained. The duration of the research (3 years, 2014-2016) and its three phases (Pilot, Study 1, Study 2), as well as the research instruments are described, all accompanied with supporting evidence. In the second subchapter the research questions are stated, followed by detailed information on the participants, research apparatus, research and data collection procedures, frequently completed by tables and charts for more transparent illustration and understanding. The chapter is concluded with limitations of the research.

The fifth chapter **Data Analysis** presents all the data accumulated, with detailed description and further comment as regards the individual phases of the research project. It is divided into sections detailing *Quantitative and Qualitative Analyses*. The quantitative analysis concerns the test data collected and makes use of descriptive statistics and statistical modelling in order to assess the comparability of the two testing modes. The qualitative analysis seeks to explore the advantages and disadvantages of the PPT and CBT modes as viewed by students.

In the sixth chapter, **Merging the Data: Findings and Discussion**, the quantitative and qualitative data are integrated in order to answer the research questions, for example, whether the preferences for the two modes of administration affect the students' test scores gained.

Finally, in **Conclusion**, the most fundamental findings are presented and further research in the area is suggested.

2 FUNDAMENTAL CONCEPTS IN LANGUAGE TESTING

2.1 A Brief Historical Context

In this chapter, Spolsky's (1977) categorization of the history of language testing into three main eras, namely Pre-scientific, Psychometric-structuralist and Psycholinguistic-socio-linguistic, which Morrow (1979) rather poetically calls the Garden of Eden, the Vale of Tears and the Promised Land respectively is introduced and discussed. Brown (2005) uses Hinofotis' term Integrative-sociolinguistic instead of Psycholinguistic-socio-linguistic but Spolsky's terminology is adhered to in this dissertation. The approaches of Communicative Language Testing, dealt with here separately even though Morrow (1979) considers it a part of the Psycholinguistic-socio-linguistic phase, Formative Testing and Assessment for Learning, are also examined. Some teaching methods are mentioned too, since testing methods tend to follow teaching methods. As Brown (2010) confirms, 'language testing trends and practices have followed the shifting sands of teaching methodology' (p.12). The expression 'follow' used in both of the sentences is of crucial importance here as it is vital to realize and will be later demonstrated that changes in testing practices take a considerable amount of time and do not happen simultaneously with the changes in the language classroom.

2.1.1 Pre-scientific era

The first era is referred to as Pre-scientific or Traditional because a well-established theory for language testing simply did not exist. Farhady (2006) claims that language ability was tested 'through subjective measures such as translation tasks and essay-type question sets' (p.33). These techniques originated in the so-called Classical or Grammar Translation teaching method, which, as its name suggests, was heavily reliant on translation and grammar. The evaluation was very subjective (Farhady, 2006) and neither attempts at objective scoring nor other aspects of language testing considered crucial today, such as validity, reliability, practicality or fairness were of much concern.

2.1.2 Psychometric-structuralist era

Carroll's paper presented at a conference in Washington in 1961, which draws attention to Lado's pioneer book on Language Testing (1961), is often considered to mark the beginning of a more professional and scientific outlook on language testing. The designation of this second era as mentioned above comes from Spolsky's (1977) term Psychometric-structuralist approach, which is heavily influenced by behaviourist psychology and structural linguistics. The main teaching method at the time is Audiolingualism, stemming from behaviourism and thus making use of repetition, positive reinforcement and avoidance of mistakes at all costs. That goes hand in hand with structuralism and its basic tenet of breaking down language into sentence-level grammatical paradigms. Test developers follow that trend and construct objective tests measuring one trait at a time, the so-called discrete point tests, which will be further discussed in 2.3.2. Favoured techniques include multiple choice tests of grammar, vocabulary and phonetics, in which reliability, a term, which will be presented in 2.2.2, comes into the limelight. As Morrow (2012) asserts the psychometric-structuralist approach 'developed the idea of language testing as measurement' (p.142). For more details about the principles of test construction following this approach see, for example, Lado (1961) and Valette (1967). Morrow (1979) characterizes Lado's approach of breaking down language complexities into isolated segments as atomistic and argues that though this form of test construction brings about easily quantifiable data, the question of how relevant and representative such data is remains unanswered.

2.1.3 Psycholinguistic-socio-linguistic era

The third Psycholinguistic-socio-linguistic phase reflects the developments in psychology and linguistics once again and thus the relevance and suitability of behaviourist psychology and structural linguistics in the language teaching and testing context are brought into question. With the advent of cognitive psychology and Chomsky's generative-transformational theory, a new approach in language testing called the Integrative theory is introduced (Farhady, 2006). The main teaching method associated with this phase was Terrell's and Krashen's Natural

Approach, which aims to develop communicative skills with the help of comprehensible input provided in a low-anxiety context. The test developers made use of cloze tests and dictations as the main test techniques and one can notice a discrepancy between the teaching and testing approaches adopted. The author of the dissertation believes that measuring communicative skills by means of dictation does not seem to go together well as dictation is not a direct form of testing speaking skills. In addition, due to it being rather unpopular with students because of the stress involved, the low anxiety environment condition is not fulfilled either. However, by moving away from discrete items, it is believed that integrative tests, which simultaneously engage multiple elements of language, could provide a more holistic picture of test takers' language abilities. Oller (1979) sees a great potential in cloze tests because he believes that they require contextual knowledge beyond the sentence. It is the exact same reason why Carroll is previously very critical of cloze techniques (1959). Contrary to Cronbach and Meehl's (1955) multi-trait approach, which asserts that different ability tests should yield different results, Oller (1979) claims that results of various abilities tested are often very similar. Brown (2010) states that the proponents of this theory form the unitary trait hypothesis, which stands on the grounds of the indivisibility of language proficiency, emphasizing that 'discrete points of language could not be disentangled from each other in language performance' (p.14). This hypothesis is later abandoned following a series of debates. (Farhady, 1982, Oller, 1983) Yet, integrative tests are still widely used primarily because of their practicality, despite the fact that further research shows that they do not assess much more than knowledge of grammar and vocabulary. For example, Weir (1990) criticizes integrative tests for focusing on test takers' linguistic competence but not the actual performance.

2.1.4 Communicative language testing

The following fourth phase reflects the Communicative Language Teaching method and is entitled Communicative language testing. It places emphasis on real life tasks and test takers' performance and it also brings the role of social context to the fore. Brown (2010) claims that tasks before communicative language testing are

'artificial, contrived and unlikely to mirror language use in real life' (p.14). According to Green (2014), the aim of communicative language testing is 'not to test knowledge of language systems, whether as discrete components or integrated whole, but the ability to use language functionally to carry out real world tasks' (p.199). It thus contradistincts the previous focus on grammar testing. Morrow (1979) provides a detailed list of characteristics of communicative language tests in his influential and laudatory article entitled '*Communicative Language Testing: Revolution or evolution*'. He states that such tests should assess test takers' performance of authentic language tasks with the help of criterion-referenced measurement (see 2.3.1 on criterion-referenced tests for more details). Such tests should establish their own validity (a term discussed in 2.2.1) and should be quality rather than quantity oriented. Reliability and objectivity, while still considered important, are only secondary to validity (Morrow, 1979). Typical tasks include task-based tests, problem solving tasks, role-plays and simulations and require test takers to use language in order to fulfil explicit purposes. The emphasis is thus placed on a direct way of testing language abilities with the help of tasks that are relevant and appropriate for each individual test taker.

Communicative tests have been criticized for not clearly focusing on testing language abilities. Stevenson (1985) notably voices his reservations concerning communicative testing in stating that a language test is 'a test not a tea party'. Bachman (1990) also criticizes communicative emphasis in testing for being a substitute for a clear definition of knowledge and Alderson (1988) claims that communicative testing implies a personal language test for every test taker since no two learners have identical needs, which makes communicative tests highly impractical and unrealistic. Furthermore, in 2012 more than 30 years after the publication of his article referred to above, Morrow, despite originally being one of the main proponents of communicative language testing, states that the most important thing about communicative language testing is that it belongs to history (Morrow, 2012). This is a thought-provoking statement, which once again points to the discrepancy between the teaching and testing spheres. Morrow (2012) adds that while communicative language teaching 'has become the default position for

language teachers in classrooms all over the world, very few testers would claim to adopt an overtly communicative perspective' (p. 140).

2.1.5 Formative testing

Green (2014) adds two more phases to the outline above, namely Formative testing and Assessment for learning. It should be noted that Formative testing first comes about already in the 1960s long before Communicative Language testing, so historical linearity is not adhered to here. It is very often discussed together with Assessment for learning (see Bennett, 2011, Black and Wiliam, 2012), hence the categorization. Formative testing has been of great importance since the Mastery Learning movement as it makes testing a crucial part of the language classroom. The Mastery Learning movement is centred around Bloom and his belief that learners need to achieve a level of mastery before moving on to the next stage. They are tested and if their results do not reach the mastery level, they are given support and then tested again. (Bloom, 1968) The idea of sequencing instruction is far from new but the idea of remedy and helping those who do not succeed is very innovative at that time. Furthermore, everybody is previously expected to go through the same amount of material and progress at the same speed and what Mastery Learning does is bring about the idea of streaming and thus greater variety (Green, 2014). Learning aptitude is considered vital in this respect and clearly defined behavioural objectives come into existence to help monitor the learners' progress. Formative testing will be contrasted with Summative testing and discussed in more detail in 2.3.3.

2.1.6 Assessment for learning

The last phase as proposed by Green (2014) is referred to as Assessment for learning and moves from the emphasis on the product of learning to the emphasis on the process of learning. It represents a more recent constructivist approach and covers a wide arsenal of learner-centred assessment practices. Gardner (2012) posits that for this approach 'assessment is our focus but learning is the goal', which summarizes the shift from testing as the necessary evil to assessment as the means to help students get better (p.2). The basic tenets of this approach are clarified with

the help of the ten principles of Assessment for learning formulated by the Assessment Reform Group (ARG). According to them, Assessment for learning:

- is part of effective planning
- focuses on how students learn
- is central to classroom practice
- is a key professional skill
- is sensitive and constructed
- fosters motivation
- promotes understanding of goals and criteria
- helps learners know how to improve
- develops the capacity for self-assessment
- recognizes all educational achievement

(ARG, 2002 qtd. in Gardner, 2012, p.3)

Black and Wiliam (2012) map out four main areas of Assessment for learning, namely classroom dialogue, feedback through marking, peer and self-assessment and formative use of summative tests (p.19). The last area mentioned overlaps to a certain degree with the Formative testing phase depicted above. The whole Assessment for learning actually stems from the formative testing movement but differs from it in the following way. At the core of formative testing of the 1960s is the interventionist approach, which Poehner (2008) describes as a static form of assessment making use of standardized test procedures and forms of assistance in order to gain easily quantifiable results, which enable comparisons between groups of test takers. One can detect the underlying psychometric approach principles quite easily. Assessment for learning, in contrast, would be better suited to the interactionist approach, a term proposed by Lantolf and Poehner (2004), which is more of a dynamic assessment and views the individual learner's needs as a priority. Poehner is one of the proponents of Dynamic assessment, which is a form of a learner-centred assessment intertwined with Vygotsky's Zone of proximal development and scaffolding (for more details see for example Kozulin & Vygotsky,

1986 or Poehner, 2008). One can clearly see the parallels to Assessment for learning. The word testing has been abandoned here on purpose since this approach makes use of other assessment means than tests only. Typical forms include self-assessment checklists, portfolios and learner logs (Green, 2014, p. 207). Bennett (2011) equals the term Assessment for learning with formative assessment and claims that it is a 'work-in-progress', which will require fundamental reconsideration of our educational ideas (p.21). Hayward (2012) adds that Assessment for learning is 'a vehicle…for sociocultural transformation where learning becomes much more of a community endeavour' (p.126). This implies a shift in the teacher-tester-student-test taker power relationship.

2.2 Test Qualities

2.2.1 Validity

Validity covers an extensive space, and all its numerous individual aspects play an unquestionable role in various areas of language tests. These aspects are usually divided into so-called internal and external, the former being content, face, and response validities and the latter referring to criterion-related validities. Then there is one of the most complex concepts, which is construct validity, the notion in particular crucial for the process of developing language tests. And finally, scoring validity, most closely related and sometimes even equated with reliability, undoubtedly embraces different features of reliability.

As discussed in the author's master thesis (p.24), one of the first attempts to answer the question whether a certain test measures 'what it purports to measure', i.e. to ascertain its validity, dates from the year 1927, and was found quite problematic (Kelly qtd. in Weir, 2005, p.12). Some sixty years later, the discussion did not appear to further develop very much, as for example, Henning's (1987) definition of validity was 'the appropriateness of a given test as a measure of what it is purported to measure' (p.89). However, the whole discourse stressed the importance of the validation process (Hughes et al., 1988), and criticism towards test misuse (when using a test designed to cover a certain area for different

purposes or a test with no specific intentions, which results in unknown validity) was regarded as serious (Alderson et al., 1995).

Although many scholars (in accordance with Alderson et al., 1995 or McNamara, 2000) thus agree to the necessity of establishing and documenting validity of use for a certain purpose, some contrasting views concerning the unifying concept of validity may be noted. Messick maintains that individual aspects of validity represent a single, unitary concept and that 'viewing different approaches to validation as separate lines of evidence for supporting given score interpretations is inadequate' (qtd. in Bachman, 1990, p.241). For Weir (2005), on the other hand, 'validity is multifaceted and different types of evidence are needed to support any claims for the validity of scores on a test' (p.13). Alderson (1995) is then another proponent of this multifaceted notion of validity, claiming that 'the more different validity aspects can be established and the more evidence gained from them, the better' (p.171). The present author adopts here, as well as in her master thesis, the unitary concept of the notion of validity, however, the individual aspects of validity will be duly described and commented on below.

2.2.1.1 Content validity

According to Alderson et al. (1995), content validity, also termed context validity by Weir (2005), affects 'studies of the perceived content and its perceived effects' (p.171). Detailed specifications of what is to be tested are recommended and experts in the field should therefore be consulted. The comparison between these specifications and test content is seen as essential in order to determine content validity (Hughes, 1989). Within this type of validity, content relevance and content coverage are distinguished by Bachman (1990). He supports Messick's division of content relevance, however, Bachman emphasizes the importance of investigating both of the aspects, ability domain and the test method facets (p.244). Many scholars, for example, Popham (1978) and Hambleton (1984), advocates of criterion referenced test development, favour this approach (Bachman, 1990). Concerning content coverage, Bachman (1990) defines it as 'the extent to which the tasks required in the test adequately represent the behavioural domain in question'

(p.245). Weir (2005) acknowledges Anastasi's (1988) guidelines, which consist in a systematic analysis of the behaviour domain and precise specifications before the test development, resulting in a suitable content coverage and its relevance.

As Bachman (1990) claims, pinpointing exact specifications for a test presents difficulties, which is mentioned by Weir (2005) as well, however, Bachman admits that in certain areas, e.g. vocabulary, 'the domain of language ability can be specified with greater precision' (p.246).

2.2.1.2 Face validity

Alderson et al. (1995) describe face validity, sometimes considered unscientific and negligible, as an internal validity component. As discussed in detail in the author's master thesis, face validity should deserve attention as it concerns the test's 'surface credibility or public acceptability' (Ingram, 1977, p.18), however, one has to admit that the face validation procedure involves non-scholars, e.g. students or administrators, whose comments are intuitive. Bachman (1990) considers face validity to be a dead concept still alive. On the other hand, some language testing experts find face validity significant, as it contributes to the test being 'accepted by candidates, teachers, education authorities or employers' (Hughes 1989, p.27). Alderson et al. (1995) agree that acceptability to users plays a significant role: 'tests that do not appear to be valid to users may not be taken seriously' (p.173). Both claim that face validity affects users' responses and is therefore closely related to response validity (Alderson et al., 1995 and Hughes, 1989).

2.2.1.3 Response validity

Response validity is concerned with the testees' responses, i.e. their approach, their procedures, and their reasoning that they employ when answering test items. According to Alderson (1995), such data gathered serve as 'important indications of what the test is testing, at least for those individuals' (p.176). The issue of response formats forms a principal part of the debate here, as 'the choice you make about format will critically affect the cognitive processing that the task will elicit' (Weir, 2005, p.62). Alderson et al. (1995) further discuss response

introspection which may reveal weaknesses in certain test items, and present another area of response validity, the process of gathering the introspective data, performed at best immediately after the test, by means of interviews or questionnaires. Despite various demerits (e.g. testees may not remember or are unable to give the reason why they produced a certain answer), it is believed that introspection examining response validity can prove valuable as regards useful insights into the issue of testing. This approach is adopted by the author of the dissertation who comments on the introspective data gathered from the testees examined in the empirical part of this dissertation (see 5.4 - 5.6).

2.2.1.4 Criterion-related validity aspects

These aspects fall within external validity, as external measures need to be incorporated; i.e. some independent assessment is suggested to be compared with the results from a given test to determine how these correlate. To establish adequately valid external criterion measures may, however, prove rather complicated, as voiced by, for example, Oller (1979) or Bachman (1990). Within criterion-related validity further aspects are recognized, which are concurrent validity and predictive validity. In the former, criterion behaviour takes place at best at the same time with the test in question, while concerning the latter, criterion behaviour is predicted for future reference. There are, of course, many limitations pointed out by, for example, Hughes (1989), Alderson et al. (1995), or Bachman (1990) and Weir (2005), as discussed in the author's master thesis (pp.32, 33).

2.2.1.5 Construct validity

Construct validity remains one of the most important, complicated as well as complex concepts in language testing. Alderson approves Ebel and Frisbie's (1991) definition: 'Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure' (qtd. in Alderson et al., 1995, p.183). According to Messick, construct validity is 'indeed the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships' (qtd. in Bachman, 1990, p.256). For

Hughes (1989), a test is said to have construct validity 'if it can be demonstrated that it measures just the ability which it is supposed to measure' (p.26). Weir (2005) emphasizes an a priori validation: 'the more fully we are able to describe the construct we are attempting to measure at the a priori stage, the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test' (p.18). The discourse revolves around the most significant question, what the construct is.

Numerous different definitions, stated in the author's master thesis (pp.34 - 39), demonstrate the difficulty of analysing the term and this creates a certain degree of obscurity. In this respect, Bachman (1990) introduces the role of various hypotheses and counterhypotheses, claiming that 'in conducting construct validity, we are empirically testing hypothesized relationships between test scores and abilities' (p.256). He favours 'logical analysis' and 'empirical investigation', and for him, construct validation is 'a special case of verifying, or falsifying, a scientific theory, and just as a theory can never be proven, the validity of any given test […] is always subject to falsification' (Bachman, 1990, p.256). Alderson et al. (1995) propose another possibility for examining construct validity, establishing four methods of gaining evidence: 1. internal correlations (i.e. correlations between individual subtests/ components and the whole test); 2. comparisons of the test results with individual students' biodata and psychological characteristics (age, gender, the first language, duration of learning, etc.); 3. multitrait-multimethod analysis and convergent-divergent validation (presented in less detail and found too complicated by him); and 4. factor analysis (reducing a complicated matrix of correlation coefficients to "more manageable proportions by statistical means"). These methods are similar to Bachman's list of forms of empirical evidence to support construct validity, however, Bachman's division, based on Messick's, provides greater detail (Bachman, 1990, pp. 271-279).

2.2.2 Reliability

Reliability also provides a useful background for the debate in the empirical section (5.1-5.3), in which further details and evidence will be discussed.

Similarly to the author's master thesis (p.39) and due to the 'growing consensus', reliability is addressed here in terms of scoring validity, which comprises different aspects of reliability, and is thus considered 'a valuable part of a test's overall validity' (Weir, 2005, p.22), corresponding to Chapelle's (1999) view that 'reliability is now increasingly seen as a type of validity evidence' (p.258). Bachman and Palmer (1996) favour 'consistency of measurement' (p.19), which is supported by Jones' claim that 'a reliable test can be depended on to produce very similar results in repeated uses" (qtd. in Weir, 2005, p.22). Weir further develops the definition, stating that 'scoring validity concerns the extent to which test results are stable over time…' (p.23). According to the American Psychological Association, 'Reliability is a quality of test scores, and a perfectly reliable score, or measure, would be one which is free from errors of measurement' (qtd. in Bachman, 1990, p.24).

Weir (2005) proposes four broad categories of scoring validity – 1. test-retest reliability (administering the same test twice to the same group); 2. parallel forms reliability (two different forms of the same test are administered on different occasions); 3. internal consistency ('homogeneity of test items'); and 4. marker reliability (written and oral exams marked by one or more markers) (pp. 24-30). The first category results in the reliability coefficient, which is calculated by means of correlating the scores. There are some apparent drawbacks, for example, the second administration of the same test proves to be rather problematic (Hughes, 1989). In the second type of scoring validity, alternate-form coefficients are gained, however, parallel forms face difficulties in testing the same level, skills, etc., as it is virtually impossible to 'construct two genuinely parallel tests' (Alderson et al., 1995, p.88). The third method stated above, called inter-item consistency by Alderson (1995), is probably the most commonly used (Hughes, 1989) and involves only one administration and one test. The usual procedure is that the performance of examinees on one half is statistically compared with the performance on the other half of the test. Here valuable results can be obtained providing the halves are equal.

The so-called split-half reliability is adopted when the items are ordered from easy to more complicated, which consists in the division of even items for one half, and odd items for the other. The last type of scoring validity, as recognized by Weir, concerns the consistency of the marker/s' individual ratings. The same set of criteria will result in 'a reliable set of ratings', on condition these are consistently employed by the marker/s (Bachman, 1990, p.178). Hughes (1989) outlines specific guidelines that informed test makers should follow to make a test more reliable (pp.36-41), and he also emphasizes a number of points to enhance the reliability of scoring (pp. 41-42).

2.2.3 Validity and Reliability tensions

The debate concerning the terms validity and reliability is perhaps somewhat outdated. Most scholars in the field agree that a test cannot be valid unless it is reliable (Alderson, Bachman, Hughes, etc.). However, validity is not a precondition for reliability, as Hughes (1989) points out. Weir (2005) describes two opposing groups: the reliability group, represented by, for example, Loevinger, claims that validity is seriously threatened without reliability, while some of those supporting validity aspects (e.g. Bachman, Cronbach, Feldt, Brennan, Wood, etc.), frequently focus on the limitations of reliability. According to Hughes (1989), 'there will always be some tension between reliability and validity' (p.42), however, the notion about validity reducing reliability and vice versa is no longer believed. It is Weir (2005) who observes in the twenty-first century that 'there has certainly been a modification of the polarized view' (p.23). Reliability aspects are then to be seen and examined as part of validity, and Weir proposes the term scoring validity be substituted for reliability.

2.2.4 Practicality

According to Hughes (1989), in order for a test to fulfil the criterion of practicality, it should be 'easy and cheap to construct, administer, score and interpret (Hughes, p.47). Bachman and Palmer (1996) look at it from a different angle and describe practicality as 'the matter of the extent to which the demands of the particular test specifications can be met within the limits of available resources'

(p.36). In their later publication they view it as the difference between required and available resources (Bachman and Palmer, 2010). It follows that an assessment lacking resources will not be maintainable. Bachman and Palmer (1996) present the following types of resources: human resources (i.e. test writers, test administrators, scorers, technical support, etc.), material resources (including space, equipment and materials) and time (development time and time for specific tasks such as designing, scoring, analysing, etc.). All these factors need to be considered, as the test can be practical, for example, in terms of human resources and time but completely impractical in terms of material resources.

H. Douglas Brown (2010) provides the following six characteristics of a practical test:

- It stays within budgetary limits.
- It can be completed by the test-taker within appropriate time constraints.
- It has clear directions for administration.
- It appropriately utilizes available human resources.
- It does not exceed available material resources.
- It considers the time and effort involved for both design and scoring.

(Taken from H. Douglas Brown, 2010, p.26)

These can be used as a checklist for test developers to see whether their tests are practical, although some items would seem to require further specifications.

James D. Brown (2005) discusses practical issues in connection with putting tests into place in language programs, which is significantly relevant for this dissertation. He draws attention to the issue of fairness to maximize objectivity and the above discussed issue of cost. He states that teachers often change their test preferences once they become test administrators and demonstrates this by his own example of originally detesting multiple choice questions and then actually arguing for using a machine scorable MCQ format because it was easy to administer and score. Similarly to the authors presented above, he also touches upon ease of test construction, test administration and test scoring. He claims that the ease of scoring is inversely related to the ease of construction, meaning that the easiest tests to

score are usually the most difficult to construct and vice versa. Finally, he stresses the importance of compromises in terms of practicality (Brown, 2005, pp.26-29). (For a comprehensive yet very detailed table with multiple questions under each category see Green (2014) who has summarized practical issues discussed by some of the authors mentioned above, namely Bachman and Palmer (2010), Brown (2005) and Buck (2009), p.62).

Weir (2005) refuses to include practicality in his test validation model and generally warns against thinking about a method before trait. He claims that 'practicality considerations are often allowed to intrude at too early a stage and validity is … threatened rather than enhanced as a consequence' (p.49).

2.2.5 Washback

As far as washback (also referred to as backwash) is concerned, Hughes (1989) shows how crucial he considers washback to be by introducing it as the very first concept in the first chapter of his fundamental text entitled *Testing for Language Teachers*. He calls it backwash, describes it as 'the effect of testing on teaching and learning' and divides it into beneficial, which according to Messick (1996) promotes learning, and harmful, which Messick claims to inhibit learning (Hughes, 1989, p.1, Messick, 1996, p.241). Other authors observe the same division but use the terms positive and negative washback (e.g. Davies et al., 1999, Brown, 2010). A test not reflecting the course objectives could serve as an example of harmful or negative washback. Washback is said to fall under and be a part of an umbrella term test impact, the former being linked to mostly classroom-based issues, functioning at the 'micro' level, while the latter implies institution or the so called 'macro' level (Green, 2014). Carr (2011) warns against using the term washback to refer to the overall test impact but claims that the confusion is understandable since washback is an essential part of impact. Green (2014) posits that 'washback connects the design and use of an assessment with what teachers and learners do in the classroom when preparing for that assessment' (p.86). It follows that tests should reflect the curriculum. Hargis (2003) goes as far as claiming that tests can become the curriculum, however, teaching for the test has

also been criticized and a balance needs to be aimed at in this respect. According to Carr (2011), curriculum-assessment mismatches are quite common too and inevitably result in harmful washback.

This is in accordance with Green (2014) who emphasizes that when an assessment validly measures the abilities to be acquired, teachers should teach towards that assessment, while in case of, for example, construct under-representation, both teaching for the test and subsequently the test will have negative washback. Hughes (1989) details the following steps to be taken in order to achieve beneficial washback:

Testing abilities whose development is to be encouraged

- wide and unpredictable sampling
- usage of direct testing
- making testing criterion-referenced
- basing achievement tests on objectives
- ensuring student and teacher familiarity with the test
- providing assistance to teachers
- counting the cost

(Taken and adapted from Hughes, 1989, pp.44-47)

H. Douglas Brown (2010) summarizes most of Hughes' steps under 'conditions for peak performance by the learner' but adds that tests with beneficial washback should be formative rather than summative (which can also be linked to criterion-referenced rather than norm-referenced tests – see 2.3.1 and 2.3.3 for details) and should provide learners with feedback in order to enhance their learning (p.38).

2.2.6 Providing Feedback on test performance

Ambrose et al. (2010) define feedback as 'information given to students about their performance that guides future behaviour' (p. 125). Šeďová et al. (2012) claim that feedback always communicates evaluation, which is in agreement with other authors who believe that feedback is an inherent component of assessment (e.g. Brookhart, 2008, Brown, 2005). Brookhart (2008) claims that the power of

feedback stems from its ability to engage learners' affective, cognitive as well as motivational factors. Feedback can be classified into many types but it is not the aim of the dissertation to present all of them here. Green (2014) comes up with a comprehensive list of characteristics of effective feedback based on Wiggins (1998), Wiliam (2011), and Hill and McNamara (2012). According to those authors, effective feedback should be prospective, i.e. looking forward rather than backward, directly related to learning goals, continuous, specific rather than general, frequent and corrective yet not excessive (Green, 2014, p.92). What should be emphasized here is that feedback is usually connected with informal performance-based assessment but should not be neglected in formal summative assessments either. With washback the same approach is recommended. One should not worry only about formative assessment and its sort of inherent need to provide information on subsequent teaching and learning but also about summative end of course assessments; as H. Douglas Brown (2010) testifies 'the end of every language course… is always the beginning of further pursuits, more learning…' (p.39), so the role of washback should not be undermined.

2.2.7 Practicality and Washback tensions

The reason why the concepts of practicality and washback can be viewed as contrasting is that tests with beneficial washback are not likely to be very practical and cost-effective. Hughes (1989) draws attention to the discrepancy between cost effectiveness and beneficial washback too but claims that the latter is superior to any cost issues: 'When we compare the cost of the test with the waste of effort and time on the part of teachers and students in activities quite inappropriate to their true learning goals…, we are likely to decide that we cannot afford not to introduce a test with a powerful beneficial backwash effect' (p.47). On the basis of this and the facts presented above, it might be inferred that the role of washback is far more important than the issue of practicality, however, one must not forget that the resources available make the testing possible and without them, there would be no washback at all.

2.3 Test Types Dichotomies

2.3.1 Criterion-referenced and Norm-referenced tests

Criterion-referenced tests (CRTs) assess how much of the material each student has learnt and their primary concern is to find how much of the material is known irrespective of the results of other students. Such tests are designed to measure well-defined objectives, which are specific to a certain course or programme. Hughes (1989) states that the purpose of CRTs is to determine whether the test takers are able to perform given tasks satisfactorily.

J. D. Brown (2005) emphasizes that there is no reference regarding the students' position vis-à-vis each other (p.4), so what matters is only whether the test taker as an individual has demonstrated the knowledge of the content tested or not. The scores are usually reported in percentages of correct responses.

Norm-referenced tests (NRTs), on the other hand, are designed to measure more general abilities independent of any courses taken or specific objectives mastered. According to H. Douglas Brown (2010), the purpose of NRTs is to put the test takers along a mathematical continuum in rank order, so NRTs relate test takers' results to the performance of others and such comparisons are made with the concept of normal distribution resulting in strong candidates being placed on one end of the continuum and weak candidates being placed on the other. The results are reported in the form of a numerical value and a percentile.

Bachman and Palmer (2010) present two contrasting approaches to score interpretation and call them relative and absolute. The aspect of competitiveness is emphasized in the relative score interpretation while the fulfilment of certain criteria is crucial in the absolute score interpretation. Parallels can unquestionably be drawn to the NRT-CRT dichotomy with NRTs being relative and CRTs absolute.

CRTs are based on instructional objectives and are thus most commonly used for classroom-based assessment while NRTs are designed to measure global language abilities and are more appropriate for standardized large-scale testing. In the context of the dissertation, CRTs are represented by the progress tests taken throughout the test takers' first year, which are further examined in the empirical

part and NRTs are represented by, for example, TSP (which is a part of the entrance exam procedure to universities, SAT or GRE tests in English speaking countries), entrance exam tests and various proficiency exams.

As for the structure of the tests, NRTs are usually longer, include more subtests and make use of versatile tasks. The actual language points are hard to predict as NRTs are used for 'assessing abstracted language ability traits' (Brown & Hudson, 2002). CRTs tend to be shorter, include numerous short subtests and the actual language points are predictable and known to the students. Brown (2005) argues that teaching to the test, which has often been criticized (see Phelps, 2011 for more details), should be a major part of what teachers do if the objectives of a course have been properly established. Oller (1979,) emphasized the instructional value of CRTs, which Brown (2010) covers by the term feedback. Feedback as an essential part of both the learning and the testing processes will be discussed in more detail later.

It would be wrong to assume that the above mentioned test types are mutually exclusive, however, it needs to be emphasized that they serve different functions as well as purposes (Bachman, 1990).

2.3.2 Discrete-point versus Integrative tests

The second dichotomy to be presented and discussed here was touched upon in the historical overview in 2.1.2 and concerns the content of the language tests. The first type – discrete-point, as its name suggests, measures small distinct pieces of language with the assumption that by collecting such information about the test taker and then putting it together, the final score arrived at will produce information on some global aspects of language ability (Brown, 2010). Regarding discrete-point tests, Lado (1961), according to Green (2014), merges a structuralist approach to language and psychometric testing to create a more scientific approach to assessing language. The main idea behind this is that each element should be tested separately and the tasks should not be related. This is usually done with multiple-choice questions (MCQs), which will be further discussed in 3.2.9.1. Carr (2011) presents the following advantages of discrete-point testing. The fact that the

items should be brief enables the test developer to cover a large number of points. Furthermore, if a test taker gets an item wrong, it should be caused by a lack of ability in that specific area rather than interference with some other areas that are being tested simultaneously. Finally, he believes that discrete-point tests are suitable for testing specific areas of language, such as grammar points covered in a course.

However, discrete-point tests have been criticized for not providing a complete picture of what the learners can actually do and for being decontextualized and thus not authentic (Brown, 2010, Carr, 2011 et al.). This led to the development of integrative or what Carr (2011) refers to as integrated tests. These tests require test takers to demonstrate knowledge of various aspects of language ability and do so by having the test takers carry out more authentic tasks. Brown (2005) describes integrative tests as those 'employ[ing] different channels and/or modes of the language simultaneously and in the context of extended text or discourse' (p.30). A typical example would be dictation as it tests many different aspects at the same time in both the productive and receptive modes.

2.3.3 Summative versus Formative assessment

The third dichotomy relevant to this dissertation concerns the function or purpose of assessment, which can be summative or formative (See also 2.1.5 for the role of formative assessment in the historical context). This division was first proposed by Scriven (1967) with respect to programme evaluation and then adopted by Bloom (1968) and used in the context of student assessment. Summative assessment aims to report on students' achievement while formative assessment aims to help students' learning (Harlen, 2012). Brown (2010) believes that most of classroom assessment is formative and should be informal with the 'ongoing development of the learner's language' as its primary concern (p.7). The information gained should guide subsequent teaching and learning. Brown (2010) further states that the key to success in formative assessment is 'the delivery and internalization of appropriate feedback on performance' (p.7). Carless (2007) divides formative assessment into 'pre-emptive', which is based on anticipation of student problems

and attempts to eliminate those before such problems arise and 'reactive', which happens 'after incomplete understanding has occurred' (p.176). Summative assessment, from the Latin origin summa meaning total, provides an overall summary of the attainment of the abilities tested. As the goal of summative assessment is to summarize, it happens at the end of a course or a unit. Summative assessment thus often includes high point value evaluation and comes in the form of grades, which are based on test scores (Bachman and Palmer, 2010) as opposed to formative assessment.

Criterion-referenced progress tests which test students' progress in learning after a shorter period of time, have often been linked to formative assessment, whereas achievement tests which also tend to be criterion-referenced but typically happen at the end of a course to determine whether course objectives have been met, have been considered summative. However, it should be mentioned that such tests also have a formative role since every test should provide learners with feedback. While providing feedback should be self-evident with formative assessment, this claim needs some justification with summative assessment, which is focused on reporting on rather that feeding into learning (Harlen, 2012). Here, washback, that is the impact of a test on subsequent teaching and learning, comes into play and as Harlen (2012) testifies: 'there is an obligation to ensure that assessment for summative purposes is conducted so that what is assessed and how it is assessed have a positive impact on learning' (p.88).

Bennett (2011) states that the definition of the term formative assessment remains unclear and divides opinion. Some understand it as an instrument, a kind of 'interim' assessment (the progress test view) while others, usually teachers and researchers rather than test developers, argue against seeing it as a test and view it as a process (p.6). Over time, proponents of the process view have started to use the term Assessment for learning instead of formative assessment to distinguish it from assessment of learning, i.e. summative assessment (Bennett, 2011). Bennett (2011) concludes that 'formative assessment might be best conceived as neither a test nor a process, but some thoughtful integration of process and some purposefully designed methodology or instrumentation' (p.7). Bennett is also critical of calling formative

assessment another name as he believes that this may only 'exacerbate, rather than solve, the definitional issue' (p.7).

Harlen (2012) remains doubtful about the distinct perceptions of summative and formative assessment altogether, claiming that the same results can be used in different ways. This opinion is adhered to throughout this dissertation, as the author believes that it always depends on how the test results are used. For example, the same instrument (i.e. the above discussed progress test) can be used formatively, which means it can feed into learning and help students progress or it can be used summatively and determine whether tested knowledge has been attained and the required abilities demonstrated.

2.3.4 Competence versus Performance

The final dichotomy to be discussed here is represented by the concepts of competence and performance. As Brown (2007) formulates it, the basic difference between the non-observable ability to do something, i.e. competence, and the concrete observable manifestation of such ability, i.e. performance, has been present for centuries. It was Chomsky (1965) though, who first proposes the distinction between those two concepts with respect to the theory of syntax. He claims that competence is the knowledge of the language and performance is the actual language used (p.4). Competence – the idealized capacity is thus distinguished from performance – the production of the utterances. Competence could be demonstrated by knowledge of the systems (e.g. grammar, vocabulary), while performance could be linked to skills (e.g. speaking, writing). Lyons (1996) further highlights the double meaning of performance as denoting the process and the product, and suggests it is not a dichotomy but a trichotomy.

Chomsky's competence-performance model has been criticized mainly because of the notion of the idealized hearer-speaker competence. (See Tarone, 1988 or Stubs, 1996 for more details.) Linguistic competence is originally associated with a narrow concept of grammar only and that is why Hymes (1972) develops the concept of communicative competence to broaden the concept and replace the dichotomy. Widdowson (1983) agrees that to use the language communicatively,

knowledge of the language together with the capacity to use that knowledge is necessary. This also suggests a uniting rather than a dividing perspective. As discussed in the author's master thesis (2006), the inclusion of knowledge and the ability to utilize that knowledge proves revolutionary and incites many educationalists to engage in this topic. Canale and Swain (1980) divide communicative competence into three main components – grammatical, sociolinguistic, and strategic. Canale (1983) then refines their earlier model and adds a discourse competence. Bachman (1990) follows that trend and states that 'performance on language tests is affected by a wide variety of factors and an understanding of these factors and how they affect test scores is fundamental to the development and use of language tests' (p.81). Bachman thus elaborates on the scheme and classifies communicative language ability into language competence, strategic competence and psychophysiological mechanisms. Language competence is further divided into organizational, including grammatical and textual competences, and pragmatic, including illocutionary and sociolinguistic competences. With respect to testing, strategic competence is understood to be 'a set of metacognitive strategies…providing a cognitive management function in language use' (Bachman and Palmer, 1996, p.71). The last component is represented by psychophysiological mechanisms and contains neurological and physiological processes, which come into play during the execution stage (Bachman, 1990). There have been many alterations and additions to this scheme but they will not be discussed here, as they do not relate directly to the context of the dissertation. (For more details see, for example, Hedge, 2000.)

3 Computer-Assisted Language Testing

3.1 Digital Literacy

Digital literacy, as mentioned above in the Introduction, forms another fundamental theoretical concept included in this dissertation. Computer literacy, a term preceding digital literacy, should be touched upon first. This older term comprises the knowledge and skills to use computers efficiently, placing the computing device and how to operate it at its core. Digital literacy goes further and covers the ability to use a much wider variety of digital devices, including laptops, smartphones, tablets, etc., with focus on the network rather than the operation of the device itself.

Dudeney et al. (2013) define digital literacies as 'the individual and social skills needed to effectively interpret, manage, share and create meaning in the growing range of digital communication channels' (p.2). As the definition suggests, digital literacies are considered to be a multifaceted phenomenon operating at various levels of complexity. Dudeney et al. propose a framework of digital literacies and enumerate the total of sixteen different literacy types.

1st focus: LANGUAGE	**2nd focus: INFORMATION**	**3rd focus: CONNECTIONS**	**4th focus: (RE-)DESIGN**
Print literacy Texting literacy			
Hypertext literacy	Tagging literacy		
Multimedia literacy	Search literacy Information literacy Filtering literacy	Personal literacy Network literacy Participatory literacy	
Gaming literacy Mobile literacy		Intercultural literacy	
Code literacy			Remix literacy

Table 1: Framework of Digital Literacies
(Taken from Dudeney et al., 2013, p.6)

It is necessary to have a closer look at the literacies relevant for the context of the dissertation. However, the individual descriptions of the literacies, as discussed by Dudeney et al. (2013), outside the scope of the present purposes, are not included. For the generation born after 1980, Prensky (2001) popularizes the term digital natives (which is previously used by Barlow (1996)) to refer to a population which has grown up surrounded by technology and thus been significantly influenced by digital media. As opposed to digital natives, Prensky (2001) introduces the term digital immigrants to denote those born prior to 1980. Prensky (2001) believes that digital natives process information differently from digital immigrants. In his view, digital natives are used to fast reception of information, parallel processing and multitasking, and might not have enough patience for traditional step-by-step processes (p.2). Digital immigrants, on the other hand, can usually 'speak the digital language' but still 'retain their accent' and thus although being able to use digital technologies, only turn to them as their second choice (Prensky, 2001, p.2). Prensky (2001) considers this divide to be an educational issue of primary concern, especially when students, who are believed to be digital natives, are taught by teachers, who are seen as digital immigrants, and one group fails to understand the 'language' of the other group, which in Prensky's view calls for a change in educational practices.

Prensky's metaphor of a certain digital divide between digital natives and digital immigrants has been widely adopted as well as criticized, mainly in the academia. Joiner et al. (2013) build on Prensky's distinction and update it by adding a second generation of digital natives born after 1993, the so-called Google or i-Generation. The Google generation is believed to use the Internet differently, mainly because of Web 2.0 (for definition see Dudeney et al., 2013) tools and social networking sites. Joiner et al. (2013) investigate the differences between the two digital generations but do not find any radical differences, except for some computer anxiety present with the earlier digital natives contrasted with none in the i-Generation. Computer anxiety will be discussed in more detail in 3.5.1. Palfrey and Gasser (2008) and Tapscott (1998, 2009), who coins the term net generation, can be

seen as supporters of Prensky's distinction. Bennett and Maton (2011), on the other hand, criticize Prensky due to the lack of empirical evidence and overgeneralizations.

With reference to a number of studies dealing with ICT access, usage, skills, knowledge and interest, Bennett and Maton (2011) dismantle Prensky's claims, for example, Otto et al., 2005, Livingstone and Helsper, 2007, Selwyn, 2008, Jenkins, 2009, Kennedy et al., 2009, Maton and Bennett, 2010, and Jones et al., 2010. Overall, they claim that Prensky's digital native/digital immigrant distinction and call for change in existing practices have created a certain academic form of a moral panic – a term defined by Cohen (1972) describing a state of affairs in which 'a group is portrayed as representing a challenge to the accepted norms and values in a society' (Bennett & Maton, 2011, p.173). The problem resulting in a deviancy amplification spiral (Cohen, 1972) is that a certain issue thus achieves prominence without substantial evidence, which Bennett et al. (2008) believe to be the case of digital natives.

Prensky defends his distinction in a book entitled 'Deconstructing Digital Natives' published in 2011. He claims that it was intended as a metaphor and he cannot be held accountable for how literally it has been taken (Prensky, 2011). He proposes the terms 'digital wisdom' and 'homo sapiens digital' to replace the much-disputed paradigm, however, the new labels have not satisfied the critics either (p.16). Bennett and Maton (2011) maintain that once again empirical evidence for such claims is scarce and Prensky's new terminology is vague and resembles 'an astrological forecast based on imagining' (p.179). While admitting that the younger generation, i.e. the one without a pre-digital mindset, might be more susceptible to experiment with digital technologies and potentially develop considerable expertise, Dudeney et al. (2013) claim that 'the notion of a homogeneous, digitally able generation is a myth' (p.10).

Recently, alternative terms of digital residents and digital visitors have been proposed, irrespective of the age of the population. White (2008) defines the digital resident as somebody who 'has a presence online which they are constantly developing while the visitor logs on, performs a specific task and then logs off'.

White and Le Cornu (2011) further elaborate on the typology proposed by White and stress the fact that it should be viewed as a continuum not a binary opposition. Inherently, users can place themselves anywhere on the continuum and neither end of the continuum is considered superior to the other. It all depends on a set of goals and the user's purpose in the given context (White & Le Cornu, 2011). Related to the goals, Dudeney et al. (2013) also emphasize the need to differentiate between being skilled at using technology for entertainment and being able to use technology for educational purposes. They insist that students actually require help and training with the latter (Dudeney et al., 2013).

3.2 Towards a CALT Framework

Now that a brief historical overview of language testing has been provided and the main theoretical concepts relevant to the dissertation introduced and discussed, Computer Assisted Language Testing (CALT) will be examined taking into consideration all the factors presented previously and matched to the developments in CALT.

CALT, also referred to as computer-based language testing or computer enhanced language testing is one of the Technology Enhanced Assessment (TEA) areas. Although it seems to offer a great deal of potential, discussed below, it has still not yet realized that potential nor brought about any major changes in language assessment (Carr, 2011, Douglas and Hegelheimer, 2007, Jamieson, 2005, et al.). According to Green (2014), 'the potential of computer-based test delivery to support innovation in the content of assessments is only beginning to be explored. Computer-based assessments are generally conservative and tend to reflect what was already possible in paper-based assessments' (pp.216, 217).

Noijons (1994) defines CALT as 'an integrated procedure in which language performance is elicited and assessed with the help of a computer' (p.38). He distinguishes three interrelated procedures involved in CALT – test generation, interaction with candidate, and evaluation of responses (Noijons, 1994). Nowadays, when CALT is discussed, Noijons' second category, i.e. the interaction with candidate

seems to be in the limelight. Suvorov and Hegelheimer (2014) produce a more general definition and refer to CALT as 'any test delivered via a computer or a mobile device' (p.2).

Chapelle (2010) pinpoints three main impulses for the use of technology in language testing, namely efficiency, equivalency, and innovation. In terms of efficiency, she mentions Computer-Adaptive Tests (i.e. CATs which will be discussed below) and automatic scoring. As far as equivalency is concerned, the aim of computer-based tests seems to be to demonstrate that they are equivalent to pencil and paper tests. The third category - innovation – offers the most promising territory in terms of transformation of language testing but as mentioned above, is still in its infancy.

Suvorov and Hegelheimer (2014) propose the following framework for describing CALT, which will be adhered to in this dissertation for its detailed yet comprehensive nature:

#	ATTRIBUTE	CATEGORIES
1	Directionality	Linear, adaptive, semi-adaptive testing
2	Delivery format	Computer-based, Web-based testing
3	Media density	Single medium, multimedia
4	Target skill	Discrete-point, integrated skills
5	Scoring mechanism	Human-based, exact answer matching, analysis-based scoring
6	Stakes	Low-stakes, medium-stakes, high-stakes
7	Purpose	Curriculum-related, non-curriculum-related
8	Response type	Selected response, constructed response
9	Task type	Selective, productive, interactive

Table 2: CALT Framework

(Taken and adapted from Suvorov and Hegelheimer, 2014, p.2)

Suvorov and Hegelheimer claim that the first five attributes and the interactive task type are unique to CALT and the remaining attributes are applicable

to both CALT and pencil and paper-based tests. It could be argued, however, that number 4 – the target skills 4 and number 5 - scoring mechanisms are not connected solely with CALT. Without doubt, pencil and paper-based tests are also concerned with discrete-point or integrated skills testing. And computerized scoring mechanisms were in place long before computerized tests. In the next section, the individual attributes will be discussed in greater detail, further elaborated on and complemented by research carried out by other authors in the field.

3.2.1 Directionality

As for directionality, linear tests provide all test takers with the same test items in the same order. In some linear tests, candidates can review their responses, in others they are not allowed to go back and alter their answers. The former type thus formally resembles the traditional pencil and paper format to a considerable extent. In computer-adaptive tests (CATs – not to be confused with CALT), which Dunkel (1999) considers to be technologically advanced assessment measures, the test takers are given test items that reflect their abilities. According to J. D. Brown (1997, p.46), CATs display the following characteristics:

1. test items are chosen and fitted to the individual students involved
2. the test is finished when the student's ability level has been determined
3. consequently such tests are relatively short

Suvorov and Hegelheimer (2014) agree with Brown stating that 'a computer-adaptive test requires ostensibly fewer items and less time to assess' because the test item complexity is adapted based on the test taker's performance (p.3). Correct answers result in more demanding questions, while incorrect answers subsequently lead to less demanding ones, which allows for efficient placement of the students. Computer-adaptive testing is made possible because of item response theory (IRT), in which complex algorithms are used to determine item parameters, such as item difficulty, item discrimination and guessing. Pre-testing is a crucial requirement for CATs and item independence a necessary condition (Jamieson, 2005). Alderson (1986) and Larson (1987) mention a number of CAT advantages, such as reduced test frustration (a tailored CAT means that test takers do not need to answer

questions beyond their ability level), immediate feedback, reduced testing time, fewer testing administrators, and an easy removal of faulty items. Drawbacks of CATs include high cost, the necessity of a large item bank equipped with well-calibrated items, unidimensionality (i.e. all items must measure one trait), security issues and the need for a high number of test takers for the statistics to work. Computerized semi-adaptive tests could be seen as a compromise. Such tests operate with what Winke and Fei (2008) call testlets, which are sets of items used for adaptive purposes. Despite the test being adaptive at the individual item level, which is the case of the CATs discussed above, computerized semi-adaptive tests pool items of the same type from a large item bank. Suvorov and Hegelheimer (2014) assert that the term 'semi-adaptive' is not universal and not all researchers distinguish between CATs and semi-CATs (p.3). For the purposes of this dissertation, this differentiation will be observed. Finally, Winke and Fei (2008) emphasize that both CATs and semi-CATs are typically associated with large testing organizations, which have the necessary resources to develop, analyse and operate them.

3.2.2 Delivery Format

The second attribute outlined by Suvorov and Hegelheimer (2014) is represented by delivery format, which in terms of CALT means the choice of computer-based or Web-based tests (WBTs). Computer-based tests are platform dependent and run with the help of certain programmes installed on the computer, or can possibly be administered through a CD or DVD. Web-based tests, which can be considered descendants of computer-based tests, have been gaining interest of the researchers since 2000. Web-based language tests, as defined by Roever (2001), are computer-based tests 'delivered via the World Wide Web' (p.84). They are platform independent and can theoretically fulfil the asynchrony principle, i.e. be administered anyplace and anytime. WBTs contain hypertext mark-up language (HTML) file(s) located on the server which is/are downloaded to the client. The downloaded HTML data is displayed with the help of web-browser software (e.g. Microsoft Internet Explorer, Google Chrome, etc.) and once test takers indicate their

responses, their answers are either sent back to the server or a certain type of items can even be scored clientside (Roever, 2001, p.85). Roever (2001) further distinguishes two types of WBTs, namely low-tech, which runs mostly clientside and high-tech, which is heavily dependent on the server. Advantages of WBTs – as put forward by Carr (2006) include simplified test construction, affordability, and flexibility in time and space, provided that access can be uncontrolled. Drawbacks include issues with security, self-scoring, data storage and server failure or browser incompatibility (Roever, 2001, pp.88-90).

3.2.3 Media Density

Media density, the third attribute of CALT outlined by Suvorov and Hegelheimer (2014), concerns the use and integration of various media configurations in CALT. This is a somewhat neglected area as tests tend to be seen as more authentic when, for example, a listening comprehension test includes a video, or a writing task is carried out in a realistic context (Noijons, 1994, p.41). Researchers, however, worry that the construct being tested changes, which brings about a validity threat. (For more details see 2.2.1.5) Douglas and Hegelheimer (2007) claim that 'the extent to which multimedia input actually enhances the authenticity of a language test' remains unclear (p.118). The reasons for the lack of video usage in CALT have been summarized by Jamieson (2005): 'it is expensive to develop, it requires high-end technology to transmit, and it is unclear how it affects the construct' (p.238). These reasons might serve as an explanation as to why any empirical research has been so rarely carried out to explore this area, leaving this issue thus rather under-exploited.

Using images and visuals for contextualization seems slightly less controversial yet Gruba (2000) stresses their supportive as well as possibly distracting roles. Roever (2001) mentions that audio and video files are problematic because of the file size, which results in long download times, and the plug-in requirement, without which it would not be possible for the file to be played repeatedly.

Despite all the negatives mentioned above, Green (2014) remains optimistic and claims that 'we can expect to see multimedia and game-like material in the near future' used in CALT (p.217).

3.2.4 Target Skill

The author of this dissertation, unlike Suvorov and Hegelheimer (2014), believes that the fourth attribute – target skill - is not unique to CALT, which is why it has previously been addressed in 2.3.2. Both pencil and paper-based and computer-assisted language tests can focus on a single language skill/system or a set of integrated skills. The former can be assessed with the help of discrete-point tests while for the latter integrative test types are more feasible. The truth is that integrated skills assessment tends to be performance-based and such tasks are frequently difficult to develop (Plakans, 2009). Plakans (2009) claims that such assessment is characteristic of English for Academic Purposes (EAP) and English for Specific Purposes (ESP) and often involves the testing of more skills, such as reading into writing or listening into reading into writing.

3.2.5 Scoring Mechanisms

The following attribute concerns scoring mechanisms. The two evaluation options are human raters or computers. They are both applicable to traditional pencil and paper as well as computer-administered tests. Scoring by machines has a long history, starting with the IBM model 805 used in the USA for scoring pencil and paper discrete-point tests (discussed in 2.3.2) as early as 1935. The IBM machine was developed to score millions of tests taken every year in a cheap, labour-saving and efficient way. With the advent of the psychometric approach, in which, as discussed in the historical overview (see 2.1.2 for more details), reliability and objectivity were emphasized, the machines were found very helpful in computing statistics. In terms of CALT, the computerized procedure of evaluating the responses (Noijons' (1994) third category mentioned above) thus definitely dates back the longest and has a much longer tradition than the actual testing on computers, which comes about in the mid 1980s.

As far as computerized tests are concerned, dichotomous as well as polytomous selected response tasks were typically scored by computers by means of matching exact answers while open-ended questions required a human scorer to analyse the students' responses, especially the extended ones. This is no longer necessarily the case since recent rapid advances in natural language processing (NLP) have introduced automated essay scoring into high stakes assessment, for example, Educational Testing Service's e-rater or Pearson Education's Intelligent Essay Assessor. According to Burstein and Chodorow (2010), the computer systems usually provide the test taker with holistic scoring and/or diagnostic feedback in this case. (For more details on automated essay scoring see Shermis and Burstein, 2013.) As for automated short answer grading, Mohler and Mihalcea (2009) provide a useful overview of the state of the art short answer graders, such as c-rater, Autotutor, SELSA, et al., some of which require manually crafted patterns, while others make use of automatic unsupervised techniques. Automated scoring of speech is another area under scrutiny but will not be dealt with here since it does not relate to the dissertation. A self-assessment system DIALANG should also be mentioned in terms of automated scoring as it provides students with feedback on their performance with reference to the Common European Framework of Reference (CEFR), a review of their right and wrong responses and advice on improvement (Douglas and Hegelheimer, 2007).

3.2.6 Stakes

The sixth attribute presented by Suvorov and Hegelheimer (2014) considers stakes, namely low-stakes, medium-stakes and high-stakes testing. The purpose of low stakes tests is to give students an indication of their abilities and possible areas of improvement but the tests themselves do not impact heavily on students' final grades or other educational outcomes. Clarke et al. (2003) go as far as stating that as far as the students are concerned such tests have no consequences attached to the test scores. Low-stakes testing should fulfil a formative function and includes practice tests, for example, web-based tests done online with no security measures taken. Roever (2001) claims that students do not usually cheat on this type of

assessment as 'cheating would not be in their best interest'. He also advocates the use of web-based tests in connection with low-stakes testing as the advantages of taking the tests whenever and wherever (i.e. anytime, anyplace) can be fully exploited (pp.7,8).

Medium-stakes tests are commonly used in the classroom setting. Typical examples include placement and achievement tests (discussed below). They should ideally be a combination of both formative and summative procedures. Roever (2001) describes them as having an influence on learners' lives but not having 'life-altering consequences' (p.8). Chapelle and Douglas (2006) warn against leaving medium-stakes tests freely accessible and to be done by tests takers whenever suitable, as the 'score meaning is likely to be compromised', although the stakes are not as high (p.57). These tests should thus be carried out in a supervised setting, as students might be tempted to cheat in order to succeed.

High-stakes testing has a heavy impact on the test taker's life or as Green (2014) puts it 'leads to decisions with serious consequences for the assessee' (p.24). Admission tests for universities, final exams, or internationally recognized certification exams can serve as examples of high-stakes assessment. High security measures need to be taken to minimize cheating and ensure test security and tests need to be administered under strictly uniform conditions. High-stakes testing provokes a lot of controversy, especially as the formative part of assessment seems to be neglected or completely missing and the whole point of testing is brought into question. In terms of assessment without feedback on test takers' performance, which is too often the case with high-stakes tests, Brown (2010) is sceptical of its effectiveness because of negative washback effects as discussed in 2.2.5.

3.2.7 Purpose

The seventh attribute is represented by purpose. At this point the author of this dissertation would suggest a change in the order of attributes in the framework because for her test purpose would definitely be one of the first aspects to consider. Suvorov and Hegelheimer (2014) refer to Carr (2011), who distinguishes two types of tests in this respect and these are either curriculum-related or other types, which

are non-curriculum-related. Carr (2011) believes that tests are used to make decisions and that 'it is important to think in terms of types of decisions more so than types of tests per se' as any test can deviate from its originally intended usage (p.6). He provides the following table:

Curriculum-related decisions	Other decision types
Admission	Proficiency
Placement	Screening
Diagnostic	
Progress	
Achievement	

Table 3: Test Purposes
(Taken from Carr, 2011, p.6)

Now the progress, achievement and proficiency tests require further detailed comment here, as they are relevant to the context of the dissertation.
According to Carr (2011), progress tests serve the purpose of assessing how the students have mastered course content so far, i.e. the process of assessment is ongoing (p.7). It can take the form of informal teacher observation, following the students' progress without any tests administered as well as making use of some teacher-made classroom-based assessment tools, such as quizzes or tests. The assessment in whatever form should be formative and help learner progress. The type of stakes involved would most probably be low to medium.

Achievement tests, which aim to find out whether the course objectives have been met (Brown, 2010, p.9), typically happen at the end of a period of instruction and have a more summative nature. The stakes involved could be described as medium since passing or failing such tests does have certain consequences.
Authors differ on the distinction between progress and achievement tests. For example, Hughes (1989) differentiates between progress achievement tests and final achievement tests but places both of them under the category of achievement tests. Farhady (2006) talks about achievement tests only and states that they are the most commonly used tests in education. According to Farhady (2006), 'all classroom

tests, mid-term tests, and final tests fall into this category' (p.5). Brown (2010) does not distinguish between progress and achievement tests either, he, similarly to Farhady, uses only the term achievement for both of the types and claims such tests have a diagnostic function too (i.e. to find out students' strengths and weaknesses) and their formative role should not be undermined. The same test can undoubtedly be used for various purposes and Carr (2011) emphasizes that 'the question of whether a particular test is an achievement or a progress test depends upon how it is being used' (p.7). If decisions are made about the learners, then it is an achievement test and if decisions are made about the teaching process (e.g. what to focus on, whether to slow down, revise more, etc.), then it is a progress test.

Proficiency tests are in Carr's terminology curriculum unrelated, which means that they are not developed to reflect a particular course content. They are designed to test the level of language ability of the learner irrespective of how they acquired that ability or as Hughes puts it 'regardless of any training they may have had' (Hughes, 1989, p.9). However, Hughes (1989) also mentions that there are two types of proficiency tests, one is more general, usually administered by external examining boards independent of any teaching institutions (e.g. ETS, UCLES) while the other has a more specific purpose directly related to a job or a course of study. One can see a slight discrepancy here as once related to a specific course of study, can such a test still be considered to be curriculum unrelated? Carr (2011) does not make such a distinction and states that proficiency tests usually assess 'more than one narrow aspect of language ability' (p.8). Brown (2010, p.11) supports this notion when saying that proficiency tests measure 'global competence in a language' and usually consist of grammar, vocabulary, reading, listening, speaking and writing subtests. Proficiency tests are typically high-stakes and summative.

One dichotomy discussed previously (see 2.3.1) should be brought up again and that is the distinction between criterion-referenced and norm-referenced tests. Progress and achievement tests would thus fit the criterion-referenced paradigm and proficiency tests the norm-referenced one.

3.2.8 Response Type

The penultimate attribute applicable to both pencil and paper and computerized tests is labelled response type in Suvorov and Hegelheimer's scheme (2014). According to them, there are two basic response types, one selected and one constructed. In terms of what the test takers are required to do, these are often referred to as receptive and productive response items respectively (Brown, 2005). Receptive response items are usually dichotomous (scored as right or wrong) and the test takers select a correct answer from a certain number of options. Productive response items require the test taker to produce a response in the form of linguistic output of diverse length.

Green (2014) adds two more – extended and personal response types. One could view at least the former as falling under the category of constructed responses but Green views both of them as distinct types. Brown (2005) also uses the term personal response items and describes them as items encouraging students to produce responses that hold personal meaning to them.

3.2.9 Task Types

Various authors (e.g. Brown, Carr, Hughes, Alderson) usually introduce response types presented above together with task types, which are Suvorov and Hegelheimer's last attribute. Suvorov and Hegelheimer (2014) divide task types into selective (e.g. multiple choice questions), productive (e.g. short answer tasks) and interactive, which are only possible in computerized tests (e.g. drag and drop). Before proceeding any further, the terminology should first be clarified. Bachman (1990) talks about test methods and Brown (2005) about test items when detailing individual task types. Both Alderson (2000) and Hughes (1989) use the term test techniques and Hughes provides the following definition: 'they are means of eliciting behaviour from candidates which will tell us about their language abilities' (p.59). Carr (2011) prefers the term 'task format' as in his view the word format draws attention to the shape of a task rather than the content (p.26), while Bachman and Palmer (1996) adopt the somewhat general term task types. All these terms will be used interchangeably as the author believes they describe the same concept.

Carr (2011) provides three types of classifications in terms of task types. One was already discussed above and describes the format of the response, i.e. selected or constructed. The second classification concerns items and prompts. According to Carr (2011), items are typically test questions expecting the candidate to select or produce a short answer and prompts require extended answers, i.e. prompts are likely to be used in productive skills assessment. Carr's (2011) last categorization distinguishes between passage-based and independent task types. Passage-based tasks require the test taker to process additional material in order to answer the question correctly, e.g. when testing reading comprehension. Independent tasks can be answered independently of any extra material, e.g. grammar questions.

Based on the categorizations of the authors mentioned above, the author of the dissertation has devised a comprehensive table of some commonly used task formats in the university setting.

Selected response tasks	**Limited production tasks**	**Extended production tasks**	**Other types**
Multiple Choice Question (MCQ)	Short answer questions	**Spoken:** Interview	Translation
True/ False	Gap-fill	Monologue	Portfolio
Matching	Transformation	**Written:**	
Ordering	**Deletion-based:**	Dictation	
	Cloze tests	Essay	
	C-tests	Summary	

Table 4: Task Types

Only the task formats relevant to the context of the dissertation will now be discussed and elaborated on. In addition, the advantages and limitations of each of the relevant task types will be outlined in the following section.

3.2.9.1 Selected Response Tasks

According to Draper (2009), multiple choice question (MCQ) format is 'the most technically developed and widely used' of all task types (p.285). MCQs consist of a stem (usually in the form of a statement or a question) and options (usually four, one of which is the correct answer, i.e. the key and the others are incorrect answers, i.e. distractors). MCQs have been lauded for their ease of scoring, practicality, efficiency (a lot of ground can be covered since MCQs are short and thus save time in this respect), increased reliability (a large number of items make the test more reliable), and automated scoring (without the necessity of a human rater interference). These advantages have caused the MCQs to be widely used in large-scale test administration.

Hughes (2003) describes the following difficulties with MCQs: only recognition knowledge is tested, guessing may have an impact on the scores, it is difficult to write successful items, cheating may be facilitated, the technique is restrictive in what can be tested and may have harmful backwash (pp.76-78). Brown (2005) warns against five potential pitfalls and these are giving unintentional clues, implausible distractors, needless redundancy, regular patterns of correct answers, and using 'none of the above' and 'all of the above' as options (pp.48-50). Fulcher (2010) adds that stems should not contain unknown vocabulary, each item should test just one concept, trick items should be avoided, negatives should be also be avoided, and options should be similar in length and structure (pp.172, 173).

Brown (2010) suggests that item indices, such as item facility, item discrimination and distractor efficiency be used to accept, discard or revise items. Fulcher (2010) agrees and stresses the need for distractor analysis to find out which of the distractors are functional and which do not work. Farhady and Shakery (2006) have looked into the role of the number of options in MCQs by administering parallel proficiency tests consisting of three, four or five multiple choice option items to 431 students majoring in English and the findings showed no significant differences, which supported earlier findings of the efficacy of three-option item tests (see, for example, Rodriguez, 2005, Grier, 1975). Yet, most MCQs used in high

stakes testing make use of four options. Oller (1979), though admitting that multiple choice questions were versatile and could be used in both discrete-point as well as integrative tests, was quite critical of the MCQ format and stated that 'due to complexity of the preparation of multiple choice tests, and to their lack of instructional value, they are not recommended for classroom applications' (p.258).

It is significant to point out that more than 30 years later despite all the drawbacks discussed above, the MCQ format is as Farhady and Shakery (2000) put it, 'undoubtedly one of the most widely used item formats' (p.78). The advantages therefore clearly outweigh the disadvantages otherwise multiple choice format would not be so widespread. Quite a few authors (as mentioned above – Brown, Fulcher, etc.) have started producing guidelines for writing good multiple choice items where they encourage test developers to follow certain steps in order to avoid some dangers associated with the MCQ format rather than insisting that all MCQs inherently need to be bad. This could be argued to be a more useful attitude.

True/False questions are typically written as statements and are fundamentally multiple choice items with only two options. Carr (2011) considers them easier to guess correctly, very hard to write well and he also believes that they are likely to become trick questions since the correct answer is often ambiguous. That might be the reason why McAllister and Guidice (2012) consider True/False questions to be 'one of the most unreliable forms of assessment' (p.193). They are quick and easy to score but the reliability considerably suffers because of guessing. This could be eliminated by asking the students to correct the false option in order to demonstrate their understanding or by adding a third possibility of 'cannot say' when testing students' reading or listening comprehension and the statement is neither true or neither false because it is not mentioned anywhere. However, the former brings about problems with scoring and how to score students' corrections and whether these should be given the same weight as just labelling a statement true, etc., and the latter is not always applicable. Better students usually perform worse as they lose a lot of time looking for the information, which they are unable to locate. Another possibility is to introduce minus points for incorrect answers, which will discourage students from guessing. Computerized scoring can be considered a

great advantage in this respect. Minus points might, however, cause a negative washback.

Matching is another example of a selected response type. It usually consists of two columns of information and test takers are asked to find matches between the two columns (Brown, 2005). It is a technique used to test vocabulary and students can be asked to match the concept with its synonym, antonym, a picture, etc. Brown (2005) calls the matching item a premise and the items to be matched are referred to as options and provides the following guidelines: include more options than premises (to circumvent elimination and guessing), options should be shorter than premises (to minimize reading) and options and premises should be related to one central theme (p.50).

3.2.9.2 Limited Production tasks

As for limited production tasks, both short answer and gap-fill are very common test techniques in our context, often used to test grammar and vocabulary. Gap-filling, also called fill-in-the-blank tasks, require the test takers to write a word or a phrase, which fits in the context presented around the gap/s. The context can take the form of a sentence, dialogue or a passage, in which one or more words get deleted. Short-answer tasks are slightly less restrictive as the candidates are usually given a bit more freedom when answering and are not limited by the given gap. In both cases, test takers are asked to construct a response, which eliminates guessing and shows more than when they select answers only. Furthermore, both the formats tend to be easier to write well compared to selected response items but scoring is more complicated (Carr, 2011). It takes longer and the scoring key needs to be detailed and very specific. This is especially true when items are not scored dichotomously but partial credit is given. Scoring needs to be done consistently and all raters need to follow the same procedure. Automated scoring can offer consistency and assurance that all test takers' answers will be marked in the same manner. It is also easier to create a list of acceptable or partially acceptable answers for the computer to mark against. However, as most questions do have more than one correct answer, it is likely that the test takers will come up with alternatives

that the test developers had never considered and that is why such cases should be discussed and the answer key altered and updated after each sitting. This is probably even more applicable to automated scoring as the answers that differ from the ones that had been fed to the computer as correct (even in the tiniest detail of one misspelled letter) will automatically be marked wrong, so a human rater is needed to go through the answers and check.

3.2.9.3 Effective Item Checklist

Now that the task types relevant to the dissertation have been described in turn, Brown's (2005) checklist for effective item formats applicable to most task types and both the pencil and paper and the computerized test formats will be summarized and commented on (pp.43-46). In his checklist he asks the following questions:

1. *Is the item format correctly matched to the purpose and content of the item?*

Here Brown warns against mixing modes and channels, which means that it is not advisable to use, for example, MCQs to test productive skills or ask the students to read aloud to test receptive reading skills.

2. *Is there only one correct answer?*

The question of whether each test item should have only one correct answer has been very much debated (Hughes, Bachman et al.) and the general consensus is that it should. Brown (2005) agrees with that, however, he has doubts about the concept of correctness itself and argues that 'correctness is often a matter of degrees rather than an absolute' and what one considers correct, another might not (p. 43). He is fairly critical of and questions the ethics of circumventing the issue by asking test takers to choose the best answer because that leaves the decision as to what is best on the test developer, who then has too much power. Instead, he stresses the need 'to write items for which there is clearly only one correct answer. This, in the present author's view, does not reflect recent attempts to encourage test takers to be more autonomous and creative, yet it is understood that where objective tests are required, too much freedom is probably not the best strategy.

3. *Is the item written at the students' level of proficiency?*

If not, this can cause problems both ways. If written in a language, which is beyond the test takers' ability, their wrong answers might be a result of not understanding what is required of them rather than not knowing the correct answer, which obviously impacts reliability. On the other hand, if the language of the items is too easy, test takers might have problems believing the authenticity of the test and face validity will suffer as a result. Brown advises aiming at the average ability level of a particular group.

4. *Have ambiguous terms and statements been avoided?*

According to Brown, ambiguity is undesirable and confusing unless the aim of the test item is to test ambiguity. This might be relevant for some high level students but is rather rare.

5. *Have negatives and double negatives been avoided?*

Brown advocates the omission of negatives and double negatives altogether. If negatives are the grammar point to be tested, then the negative elements should be somehow highlighted (e.g. capital letters, bold font, etc.), otherwise they should be avoided altogether.

6. *Does the item avoid giving clues that could be used in answering other items?*

Logically, if one item gives clues as to what the correct answer of the following one is, then one cannot consider answers to the second item valid or reliable.

7. *Are all parts of the item on the same page?*

Students should not be put under more pressure by a user-unfriendly layout or interface. This can be easily taken care of by careful checks before test administration. In terms of pencil and paper tests, final proofreading before printing should suffice and as for computerized tests, the test should be tried out on the computers that the test takers are going to use in order to ensure browser compatibility.

8. *Is only relevant information presented?*

Brown warns against extra time being spent on processing information that does not contribute to the test results. He once again mentions that extra information might give students unsolicited clues.

9. *Have race, gender and nationality bias been avoided?*

A biased test, i.e. disadvantaging any race, religion, nationality, etc., is uncalled for as fairness and objectivity suffer. Brown adds that 'since the potential for bias differs from situation to situation, individual teachers will have to determine what is appropriate for avoiding bias' in their particular context. (Brown, 2005, p.46)

10. *Has at least one other colleague looked over the items?*

As careful and perfectionist as one test developer might be, it is impossible to spot your own mistakes when checking the test again and again on your own. Brown agrees and also claims that language tests should be tried out by native speakers, quoting Lado (1961, p.323 qtd. in Brown, 2005, p.46) who stated that 'if the test is administered to native speakers … they should make very high marks on it or we will suspect that factors other than the basic ones of language have been introduced into the items'.

There are other authors who present various lists of effective item formats (for example, Weir, 1990, Alderson et al., 1995, Hughes, 2003 and Carr, 2011), however, Brown's (2005) checklist is the most comprehensive.

3.3 CALT Advantages and Disadvantages

As presented and discussed in the author's master thesis (pp. 53, 54), James Dean Brown (1997) explores certain benefits and drawbacks concerning the effectiveness of using computers in language testing. He further divides the benefits into two categories – human considerations and testing considerations and the drawbacks into physical and performance considerations.

BENEFITS of CALT	
Human considerations	**Testing considerations**
Students can work at their own pace	Computers are more accurate at scoring selected-response items
CALTs take less time to complete and are thus more efficient (found, for example, in	Computers prove more accurate at reporting scores

Madsen, 1991)	
CALTs are less overwhelming when questions are presented one by one	Computers can give immediate feedback and statistic data
In CATs students feel less frustrated because they work on test items appropriate for their ability level.	Computers provide enhanced ability-level estimates (because of IRT and CATs)
Students like computers and enjoy the testing process when computerized (Stevenson and Gross, 1991)	Computers are efficient at providing diagnostic feedback
	Cheating should be minimized by large item pools

CALT = Computer Assisted Language Tests, CAT = Computer Adaptive Tests

Table 5: Benefits of CALT

(Taken and adapted from J. D. Brown, 1997, p.47)

DRAWBACKS of CALT	
Physical considerations	**Performance consideration**
Insufficient availability of computer equipment and/or their breakdowns	Different results if the same test is presented on a computer or paper (Henning, 1991)
Screen capacity and screen size limitations may cause problems	Varied degrees of computer familiarity may lead to discrepancies in performance
Graphics capabilities may be restricted	Computer anxiety may deteriorate test performance

Table 6: Drawbacks of CALT

(Taken and adapted from J. D. Brown, 1997, p.48)

The two tables above show that already in 1997 the benefits outnumber the drawbacks. Furthermore, bearing in mind that technology has considerably developed since then, one could probably leave out J. D. Brown's concerns relating to screen capacity and graphics capabilities. As for insufficient availability of computer equipment, this might still be an issue at schools, universities and other

institutions in our context but not so much in students' homes as, according to the data provided by the Czech Statistical Office (2013), 92.3% of Czech households with children own a computer. This should also lead to a decrease in test performance discrepancies caused by varied degrees of computer familiarity, which will be discussed below. The problems with computers breaking down and other technical glitches that even the most careful planning and technical help cannot always eliminate are still an issue and a possible reason for both test taker and test administrator frustration. The benefits in relation to testing considerations are hardly disputable since computers are faster and more efficient than human raters and in terms of automated scoring of selected response items also error free, provided the correct answers have been correctly inserted into the scoring key. As for the benefits in terms of human considerations, one could argue that questions are not always presented one by one (except for CATs) and tests are very often timed, so the students cannot work at their own pace. They need to adhere to the time limit, especially in medium or high-stakes testing situations.

Stevenson and Gross' (1991 qtd. in Brown, 1997, p.47) claim that students like computers and thus enjoy the testing process more when it is computerized can still be considered valid and applicable today, although probably not universally. There is a considerable difference between trying out low-stakes tests in the safety of one's home and experiencing a real medium or high-stakes testing situation, in which anxiety and feelings of nervousness may emerge. There are also students who like computers but do not enjoy the computerized testing process at all as the empirical part will demonstrate.

The Master thesis of the author of this dissertation presents five test method differences put forward by Chapelle and Douglas (2006), who draw on Bachman and Palmer's (1996, pp. 49, 50) task characteristics. They are examined here with respect to the two modes. The first test method difference - physical and temporal circumstances or setting in Bachman and Palmer's terminology - includes the place, the time and the people involved in administering the test. As already mentioned, computerized tests, especially web-based ones, enable a greater variety of locations, flexibility in terms of when test takers take the test and a very limited or no need of

personnel to administer the test. However, in high-stakes testing situations, this becomes an issue and might compromise security (Chapelle and Douglas, 2006, pp.25, 26). The second test method difference is represented by rubric and instructions. These tend to be consistent and presented in a uniform way, which according to Chapelle and Douglas leads to enhanced fairness (p.23). Chapelle and Douglas are critical of optional help screens and different languages of instructions, which can devalue uniformity (p.23). In contrast, it sometimes believed that instructions provided in the test takers' mother tongue lead to enhanced test reliability. Input and expected response, the third test method difference, covers the material the test taker is exposed to and the way that they respond to it. Technology offers diverse input yet it depends on the multimedia capabilities available (Chapelle and Douglas, 2006, pp.28, 29). Interaction between the Input and Response constitutes the fourth test method difference and some types of tests offer the candidates immediate feedback or, in the case of computerized adaptive testing, items tailored to test takers' abilities. The last test method difference is absent in Bachman and Palmer's scheme and Chapelle and Douglas refer to it as the characteristics of assessment including the definition of construct, criteria for correctness and scoring procedures. Here, Natural Language Processing (NLP) seems to be an underlying technique, which enables automated scoring of complex responses, however, it affects the definition of the construct as well as the scoring criteria, which can lead to potential problems (p.23).

Noyes and Garland (2008, pp.1368-1370) present some pros of online assessment in general, which are also relevant to the context of the dissertation, such as standardization of test environment, online scoring and the richness of interface, while others, for example, diverse user population or quality and quantity of composition, demonstrate much less significance to our purposes. As for online assessment cons, Noyes and Garland, similarly to J. D. Brown (1997), mention problems with computer hardware and software, namely crashing and freezing and the need to restart the computers when something goes wrong. As a result, computerized tests can take longer than pencil and paper-based ones as supported by the results of a study carried out by Zandvliet and Farragher (1997). On the other

hand, Rabinowitz and Brandt (2001), for example, find that when large numbers of test takers are involved, computer-based tests take a shorter time to administer. Another problematic area is serial presentation, here Noyes and Garland (2008) insist that 'it is easier to look through items and move backwards and forwards when using paper' (p.1369). They also reintroduce the issue of the computer screen. However, as opposed to Brown (1997), who is concerned with its capacity and size, Noyes and Garland refer to Ziefle (1998), who maintains that it can be more tiring to work on the computer than on paper. The last two cons that they discuss, namely lack of a controlled environment and concerns about confidentiality (especially in relation to web surveys and social desirability), are more characteristic of other forms of online assessment not dealt with here.

H. Douglas Brown (2010) proposes the following more specific and up to date list of advantages and disadvantages of CALT. Some similarities to J. D. Brown's list presented above are visible at first sight but some points differ considerably. According to H. Douglas Brown, computer-based testing offers:

- a variety of easily administered classroom-based tests
- self-directed testing on various aspects of a language
- practice for upcoming high-stakes standardized tests
- some individualization, in the case of CAT
- large-scale standardized tests that can be administered easily to thousands of test-takers at many different stations, then scored electronically for rapid reporting results
- improved (but imperfect technology for automated essay evaluation and speech recognition

  (H. Douglas Brown, 2010, p.20)

The CALT disadvantages H. Douglas Brown (2010) mentions include:

- lack of security and the possibility of cheating are inherent in unsupervised computerized tests
- occasional 'homegrown' quizzes that appear online may be mistaken for validated assessments

- the preferred MCQ format contains the potential for flawed item design
- open-ended responses are less likely to appear due to a/ the expense and potential unreliability of human scoring or b/ the complexity of recognition software for automated scoring
- the human interactive element is absent
- validation issues stemming from test-takers approaching tasks as test tasks rather than as real-world language use

(H. Douglas Brown, 2010, p.20)

One can see that the advantages and disadvantages described above are more concerned with the content of the tests rather than focusing on some of the technicalities of the computerized testing process as in the case of J. D. Brown's list. One of the CALT benefits proposed by J. D. Brown is actually considered a disadvantage by H. Douglas Brown, who believes that cheating may be facilitated due to the unsupervised nature of the setting. That is certainly true, however, what J. D. Brown most likely refers to is the elimination of cheating by giving different test items to different test takers in a supervised setting, which is made easier by large item banks. Further support for this interpretation can be found in the study conducted by Bodmann and Robinson (2004) who maintain that one of the CALT advantages lies in easier manipulation of test items to reduce cheating.

Similarly to H. Douglas Brown (2010), Carr (2011, p.185) also highlights improved test delivery, scoring and efficiency in comparison to pencil and paper-based tests and adds that CALT offers more attractive-looking tests and more engaging tasks. This goes against H. Douglas Brown's statement that test takers approach CALT tasks as test tasks instead of viewing them as authentic language use. Davey (2011) classifies CALT advantages into two main categories – improved measurement precision and efficiency and increased convenience. The former primarily concerns CAT models discussed in 3.2.1. The latter contains the following aspects: self-proctoring, immediate scoring, integrated data management systems, diagnostic assessment and integration with instructional software, flexible

scheduling, reach and speed and student preference (Davey, 2011, pp.2-5). Some of the aspects remain questionable as, for example, self-proctoring in medium or high-stakes test situations is not feasible even when test takers have different versions of the test. Furthermore, technical help might be required during the process, which the invigilator cannot always guarantee. As for the very much discussed advantage of immediate scoring, this is only true about selected-response items while productive response items corrections often need to be carried out by human raters. Davey (2011) suggests a compromise in which the computer 'reports what it can and full results follow after human ratings are produced' (p.4), which is the system used in the context of the dissertation. In the case of low-stakes CBTs, scheduling is definitely more flexible, however, once the tests need to be administered in an invigilated setting, booking computer rooms, etc. can prove rather limiting in terms of flexibility.

3.3.1 CALT Validity Threats

Chapelle and Douglas (2006) adopt a different perspective and identify a number of potential validity threats with respect to CALT, offering suggestions to diminish them. The summarized table presented below was first introduced in the author's master thesis (pp.49, 50).

Potential threat to validity	Further explanation	Suggestion
Different test performance	Computerized tests may not test the same abilities as those measured by other forms of assessment	Research comparing performance on parallel forms is needed
New task types	Task types characteristic of computer administration differ from pencil and paper tasks	Qualitative and quantitative examinations of the performance on the new task types need to be carried out to grant appropriateness
Limitations due to	Sampling of the items may	Experimentation with

adaptive item selection	not reflect the test content appropriately	variation in item presentation and control
Inaccurate automatic response scoring	Automatic response scoring may not assign credit to the qualities of a response	Research aimed at developing relevant criteria for evaluating machine scoring is vital
Compromised test security	CALT may present risks to test security	Issues of security need to correspond to the particular test purpose
Negative consequences	CALT may induce negative impact on learners, learning, classes, and society	Documenting, understanding and planning for the negative consequences remains a necessity

Table 7: CALT Validity Threats
(Taken and adapted from Chapelle and Douglas, 2006, pp. 41-61)

Some of the areas will be referred to here only briefly, while others, for example, new task types or possible reasons for different test performance with respect to comparability studies and individual differences will be discussed below in more detail.

As mentioned in the author's master thesis (p.51), limitations due to adaptive item selection involve concerns related to the content tested and the selection left up to the computer. Canale describes the CAT environment as potentially 'trivializing, compromising, and reductionist' (qtd. in Chapelle and Douglas, 2006, p.50).

Inaccurate automatic response scoring poses yet another threat, especially when test takers produce linguistic responses. Automated scoring of writing and speaking has already been touched upon (see 3.2.5) as well as issues with automatic scoring of constructed short answers (see 3.2.9.2). Test security and its potential compromisation is to be presented here with respect to various test types as

summarized by Davey (2011). Linear computer-based tests are considered more secure than traditional pencil and paper-based tests in terms of students copying from one another. However, approximately the same amount of risk is involved in both modes as far as disclosure through repeated administration of the same test is concerned. Reuse should be limited and parallel test forms are needed (Davey, 2011).

Similarly, random form CB tests and multi-stage tests are also more secure than traditional pencil and paper-based tests both in terms of students' copying and unintentional disclosure. Nevertheless, they tend to be marginally more secure than fixed CB tests because test security increases with the size of the item pool. Davey (2011) emphasizes that 'pools will need regular replenishment or replacement if frequent administrations are offered' in order for the test to remain secure (p.7). And with regards to multi-stage tests, more combinations of stages and testlets result in the same effect as with large item pools (Davey, 2011). Item-adaptive tests are once again more secure than traditional tests regarding students copying and unintentional disclosure. Furthermore, 'if properly configured' they 'potentially [offer] the best protection … against item overexposure and attendant security difficulties' (Davey, 2011, p.10). The difficulties of CATs and their configuration have been covered in 3.2.1. The security of the last type of a CB test, the computerized classification test, corresponds to the other pool-based tests but as it tends to aim at 'relatively a few performance thresholds, it is somewhat more prone to overexposure' (Davey, 2011, p.11).

3.4 CALT Innovative Task Types

As for new task types, a great number of authors (see 3.3 on CALT advantages for greater detail) laud CALT for the innovation it may bring in terms of new task types. However, the reality seems to fall short of the expectations. The author of the dissertation has tried to find examples of some innovative task types used in practice but without much success. Parshall and Harmes (2014) define innovative task types as 'those items in a CBT that make use of features and functions of the computer to do things not easily done in traditional paper-and-

pencil assessment' (p.1). These features include animation, sound, video, and graphics. The innovation can be incorporated within the item format, the response, the interactivity provided by the item, the media used, the item complexity, item authenticity or the scoring method (Parshall et al., 2009).

Strain-Seymour et al. (2009) summarize the advantages of innovative task types as enabling: 'Measurement of a broader range of skills, increased authenticity, improved presentation of complex and dynamic information, reduced reading load, increased student engagement, reduced effect of successful guessing, reduced demands on working memory, allowing for more valid measurement and measurement of process skills and higher-order thinking' (p.3). They produce the following overview of tasks:

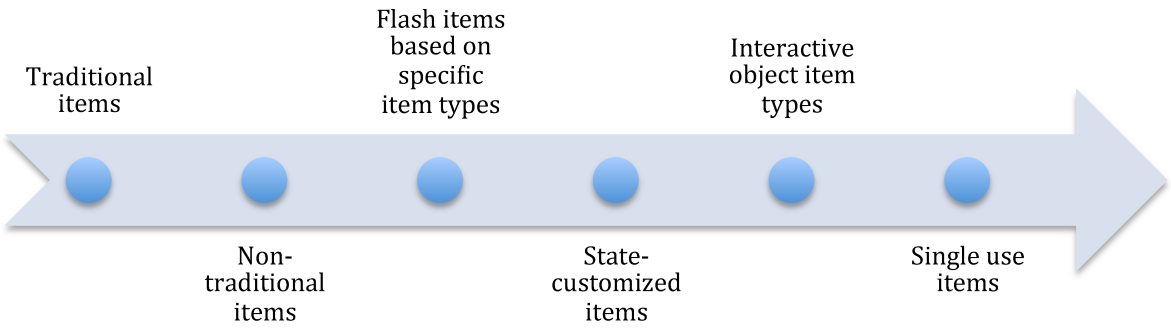

Figure 1: Continuum of Innovative Item Development
(Taken and adapted from Strain-Seymour et al., 2009, p.9)

Strain-Seymour et al. (2009) present the overview of computer-based tasks as a continuum with traditional task types with known psychometric properties and decreased cost on one end and innovative task types with increased customization but lesser known psychometric properties on the other. They consider multiple choice and gridded response items with no art, or static black and white art to be Traditional items as opposed to Non-traditional items represented by the same items but with video, animation or colour images. Examples of Flash items based on specific item types would be drag and drop, simple construction environments and graph creation. State-customized items would be extensions of existing item types

with a state-specific user experience or some customizable parameters and Interactive object item types could be characterized by combining interactivity and re-usable art (e.g. a balance beam). Finally, Single use items placed on the innovative end of the continuum could be exemplified by a non-reusable virtual lab (Strain-Seymour et al., 2009).

Ockey (2009) is one of the few authors, who actually provides some examples of innovative task types specific to language testing. He provides the following experimental question types used in DIALANG assessment of reading abilities: mapping and flowcharting, reorganization, insertion, and word deletion. Mapping and flowcharting involves two stages. Test takers first read a text and then they are asked to choose words to drag into a map or a flowchart. They are not allowed to look at the text when working with the map, which would be rather hard if not impossible to ensure in pencil and paper-based mode. Reorganization, as the name suggests, asks the test takers to reorganize sentences in the correct order to make a logical story. The author of the dissertation would argue that this particular task is undoubtedly possible in the pencil and paper-based mode too but the truth is that a computer allows for easier manipulation of the sentences and enables the test takers to view the possible orders. Insertion in Ockey's description also includes two stages. First test takers are asked to identify where a missing word should be put and second, in the case that they are correct, they can insert (i.e. type in) the missing word. This is only partially possible in pencil and paper-based mode as students can insert the missing word but without knowing if they have identified the correct place for the insertion. Word deletion forms the last experimental task type as put forward by Ockey. It resembles an error correction task, in which test takers are asked to delete words that do not fit in the sentences. The obvious advantage of the computer-based mode is that once deleted, test takers can see what the sentence looks like without the deleted word. Ockey (2009) also discusses developments in listening, speaking and writing assessments but these have been partially covered previously (3.2.5) and do not constitute the core of this dissertation, so they are not discussed here again. It is always vital to ensure that the innovative tasks 'measure the intended construct' and are not implemented only because the technology

makes it possible (Ockey, 2009, p.841). On a similar note, Douglas (2000) warns against language testing being driven by technology rather than 'technology being employed in the service of language testing' (p.275).

3.5 Comparability Studies

A great body of literature dealing with computer-based and computer adaptive tests has been produced, however, surprisingly little empirical research has focused on comparing identical versions of pencil and paper-based and computer-based language tests, although converting existing pencil and paper-based tests into computer-based ones seems to be a very common practice. Lightstone and Smith (2009) attempt to account for the scarcity of empirical research by a common belief that 'if items are identical, then the testing mode is irrelevant' (p.31) but immediately dispute that misbelief by mentioning the following studies: Bugbee and Brent (1990), Parshall and Kromrey (1993), Ployhart et al. (2003), who report better results when computer-based tests are used, while Green et al. (1984) conclude that students perform better on traditional pencil and paper-based tests and Russell (1999) shows mixed results affected mainly by keyboarding speed. Bunderson et al. (1989) in their overview of test equivalence state that scores from pencil and paper-based tests are often higher (in 13 out of 27 studies), although differences are generally small, or show no significant difference (in 11 out of 27 studies) (p.378).

Mead and Drasgow (1993) conduct a meta-analysis of 28 studies concerning the mode of delivery but also focus on two other variables – linear versus adaptive and power versus speed tests. Linear and adaptive tests were discussed in 3.2.1 but the other dichotomy needs to be clarified here. Speed tests examine how many questions a test taker answers in the allocated time, the questions being straightforward and the answers clear. Power tests, on the other hand, ask fewer questions but they are more complex and arriving at the correct – not at first sight obvious – answer poses a challenge to the test taker. Once test takers figure out the way of solving the issue, the answer itself is not necessarily complicated. In order to avoid confusion, it should be stated that both the types are timed, so there is a time

limit to observe. In their meta-analysis Mead and Drasgow (1993) conclude that pencil and paper-based tests produce slightly better results than their computerized counterparts but the only variable to significantly influence scores is speediness, which could be attributed to differences in motor skills when manipulating the speed test in the two modes. Interestingly, timed power tests do not show any mode effect on test scores. Wang et al. (2008) contradict Lightstone and Smith's (2009) claim about the lack of empirical research regarding the comparison of the different modes of test administration and state that 'test administration mode effects have been extensively studied' with more than 300 studies concerned with test mode effects on 'intelligence, aptitude, ability, vocational interest, personality, and achievement tests' (p.8). They do not, however, indicate how many of these studies deal with language tests, so Lightstone and Smith may have pronounced quite an essential statement in the argument.

Wang et al. (2008) refer to the following studies, which show that results from computer-based and pencil and paper-based tests cannot be used interchangeably: Godwin (1999), Mazzeo and Harvey (1988), and Pommerich and Burden (2000) (p.8). There are other studies, most of which relate to the field of psychology, which find no significant differences between the two formats even if the versions are not identical, such as Glowacki et al. (1995), DiLalla (1996), Ogles et al. (1998), Cronk and West (2002), Paek (2005), Williams and McCord (2006), and Žitný et al. (2012). Nevertheless, McDonald (2002) emphasizes that 'equivalence should not be assumed, but always needs to be demonstrated', which is supported by e.g. Van de Vijver and Harsweld (1994) and Brosnan (1998) (p.300). McDonald (2002) furthers his point by claiming that computer-based and pencil and paper-based tests 'are likely to coexist for the foreseeable future, with some tests existing in both formats' and he maintains that 'the issue of equivalency is therefore very significant and is probably yet to come to the fore' (p.301). Noyes and Garland (2008) admit that achieving equivalence of the two formats is problematic yet they maintain that greater equivalence is being achieved today than in the past and one can only hope that this tendency will continue.

3.5.1 Individual Differences

Fulcher (1999) insists that equivalence is not the only equity concern and that previous experience of using computers, test taker attitudes to computers and test taker background should also be considered as they can affect test scores. McDonald (2002) agrees and states that since the two formats offer different experiences to the test takers and individual differences might impact the results, it is essential to establish both score and construct equivalence, especially in settings where parallel forms of the pencil and paper-based and computer-based tests are in use. These individual differences can also cause the negative consequences of CALT on learners and their learning as outlined by Chapelle and Douglas (2006) in the table above.

Similarly to Fulcher, McDonald (2002) proposes three areas of concern, namely computer experience and familiarity, computer anxiety and computer attitudes. Weir et al. (2007) refer to Daiute (1985) who claims that in testing situations, for example, writing done with traditional tools (meaning pen and paper) produces better results than writing done on computers, which could be interpreted as potentially due to students having more experience with the former at that time. On the other hand, Russell and Haney (1997) accumulate sufficient evidence to demonstrate that test takers' writing on computers achieve better results than those produced on paper, however, in both of these cases the link between computer familiarity and better/worse results is only arbitrary and not accounted for. Some studies suggest that students who have less experience with computers have worse results on computer-based tests (e.g. Pomplum et al., 2005 and Bennett et al., 2008), while others do not confirm that (e.g. Clariana & Wallace, 2002). Taylor et al. (1998) report the small effect of computer experience on test results but after giving the test takers a computer-based tutorial, Taylor et al. (1999) conclude that 'no evidence of adverse effects on TOEFL CBT performance [is] found due to lack of prior computer experience' (p.220). Similarly, Fulcher (1999) does not find any significant impact on test results brought about by computer familiarity in his study comparing pencil and paper and computer-based versions of the English placement test. Smith and Caputi (2007) emphasize the role of practising the computerized

format before the actual test is taken, which corresponds with findings by Taylor et al. (1999) but the evidence of students consequently achieving better results is absent in Smith and Caputi's study. As far as multiple choice tests are concerned, Bennett (2002) and Bridgeman et al. (1999) come to the conclusion that individual differences in computer experience do not have a significant effect on the test scores.

In this day and age, computer familiarity and experience, at least in the context of the dissertation, should not have a great effect on the results, however, as McDonald (2002) puts it, it should still be taken into consideration when comparing pencil and paper-based and computer-based tests. The question of how to measure computer familiarity with adequate validity remains an issue and since it is measured differently in different studies mentioned above, that alone can account for the inconsistency in results.

Another factor, which might distort statistical equivalence of pencil and paper-based and computer-based tests, is represented by computer anxiety. Howard (1986, p.18) defines it as the 'fear of impending interaction with a computer that is disproportionate to the actual threat presented by the computer'. Computer anxiety entails components pertaining to behaviour (i.e. resistance towards computers), emotion (i.e. fear of computers), and attitude (i.e. aggression/ hostility to computers) (Brosnan, 1998, p.12). That might explain the fact that the terms computer anxiety and computer attitudes were originally often used interchangeably, however, Kernan and Howard (1990) stress that these need to be treated separately and their view will be adhered to in this dissertation. McDonald (2002) does, however, state that computer anxiety overlaps with computer confidence and views the two as the same construct. He provides the following studies of Levine and Donitsa-Schmid (1998) and Powers (1999) as evidence. Computer anxiety is said to stem from a lack of computer experience, which would suggest that with increased exposure, computer anxiety should be on the decrease. According to McDonald (2002), studies have shown rather conflicting findings in this respect. For example, Gos (1996) discloses that quality of exposure is just as crucial a factor as quantity in the development of computer anxiety and thus

emphasizes the need for positive experiences when working with computers. Beckers and Schmidt (2003) go even further and claim that the amount of computer experience is not the decisive factor but that positive experiences actually lead to reduction in computer anxiety. Chien (2008) observes that computer anxiety is often linked to test taker's attitude towards computers. Computer anxiety is therefore identified with a negative attitude towards computers.

According to Whitley (1997), the two concepts are interconnected but should be treated separately. Kernan and Howards (1990) agree and find evidence that students with high computer anxiety do not necessarily have negative attitudes to computers. Gender is often considered to be an element influencing computer anxiety and studies point to the fact that women are more anxious about computers than men (e.g. Brosnan, 1998, Broos, 2005). Chien (2008) points out that gender bias could be the reason behind this phenomenon and suggests that 'since boys spend more time using computers, they have more computer experience than girls' (p.20/2). He further supports his statement by studies carried out by Bannert and Arbinger (1996) and Beentjens et al. (1999), who argue that boys are simply more interested in computer activities (Chien 2008). Young (2000) investigates gender differences among high school students, and finds that computer use is considered to be a male domain. Chien (2008) presents five more aspects that affect computer anxiety, namely age, personality traits, math anxiety, and social-economic background, but these will not be elaborated on here as they are not examined in this dissertation and the studies provide rather an inconclusive picture. Hargreaves et al. (2004) draw attention to the fact that young people in the UK are familiar with using computers yet they are still used to taking pencil and paper-based tests rather than computer-based ones. According to Green (2014), 'writing with pen and paper [is] replaced by word processing' (p.216), so it is only logical that the testing situation should reflect this trend. The author of this dissertation believes that this aspect alone can lead to computer anxiety because students are simply not used to computer-based testing. Weir et al. (2007) point out that anxiety can also have positive effects on the students' results but such claims lack empirical research.

The third factor possibly having an impact on the test scores as proposed by

McDonald (2002) concerns attitudes towards computers. Attitudes to computers are often influenced by computer familiarity and computer anxiety, which, as documented above, can be considered an example of an attitude towards computers, so the concepts are closely intertwined (McDonald, 2002). Positive attitudes should be examined too, since they have been linked to increased computer use and vice versa (Levine and Donitsa-Schmidt, 1998). Nevertheless, this is not always the case as it also depends on the quality of exposure as mentioned above. Stricker et al. (2004) come to the conclusion that in their study test takers' attitudes towards the computer-based TOEFL (Test of English as a Foreign Language) are moderately positive. Surprisingly, they find that computer familiarity is 'unrelated to attitudes about the computer-based TOEFL' (p.49). Various scales to measure computer attitudes have been developed (Levine and Donitsa-Schmidt, 1998, Selwyn, 1997, Nickell and Pinto, 1986, Reece and Gable, 1982, Kay, 1993). For example Levine and Donitsa-Schmidt (1998) include five main areas in their computer attitude scale: 'computer self-confidence, attitudes towards computers as an educational tool, stereotypical attitudes, perception of computers as a tool for enjoyment, and importance of computers' (qtd. in McDonald 2002, p.307). Computer self-confidence has often been linked to self-efficacy, i.e. an individual's belief in one's capability to perform a task (Bandura, 1977). Despite the fact that Garland and Noyes (2008), who analyse four widely used attitude scales, consider those scales reliable, they emphasize the need to use and interpret them carefully if at all due to changes in the construct over time. McDonald (2002) insists that a very limited number of studies connect computer attitudes to actual test performance and those which do so show either no significant effect (Fulcher, 1999) or mixed results (Russell, 1999). Further research is necessary and should be addressed in comparative studies.

4 RESEARCH METHODOLOGY

4.1 Research Approach

The research approach adopted in this study can be characterized as complementary, combining quantitative as well as qualitative research instruments. A convergent parallel mixed methods design is made use of in order to merge different types of data collected and 'to provide a comprehensive analysis of the research problem' (Creswell, 2014, p.15). Given that the research is longitudinal and took place from 2014 to 2016, it can be considered a multi-phased mixed methods design, too, with each new stage building on what has been learned in the previous one, including the Pilot (Creswell & Clark, 2011, p.100).

A repeated-measures design, which Johnson and Christensen (2008) consider a strong experimental design, has been chosen for the quantitative strand of the study to repeatedly measure all participants under each treatment condition. Since the same participants take part in all experimental conditions within one stage, there is no need to equate participants from different groups and the participants 'serve as their own control group' and are thus perfectly matched (Johnson and Christensen, 2008, p.320). The counterbalancing technique, which will later be described in more detail, has been employed to overcome sequencing effects, namely the order and the carryover effects (Johnson and Christensen, 2008, p.303).

The qualitative strand is represented by an attempt at a grounded theory approach as the researcher collected qualitative data from the participants and did not limit their output by any existing concepts or theories. While the general advantages and disadvantages of CALT have been given a lot of attention in the language testing literature (as the theoretical part demonstrates, see 3.3), these have been mostly considered from the general perspective of test developers and therefore there is a lack of information coming directly from the students. According to Charmaz (2005), Glaser and Strauss's (1967) term grounded theory involves 'simultaneous data collection and analysis, with each informing and focusing the other throughout the research process' (p.508). Maxwell (2013) describes it as a

theory, which is 'grounded' in or inductively developed from the data collected (p.49). Repeated themes and concepts are identified in the data collected and are coded using an open emergent coding system. Based on the data gained, a questionnaire is designed and piloted and it is hoped that it might serve as a measure of students' attitudes towards computer-based testing in the future.

The main research instruments include achievement language tests administered in the computer-based and pencil and paper-based modes and self-report feedback forms administered on paper. Through the triangulation of the research instruments, it is believed that more valuable and in-depth data will be gained, in particular due to integrating all the information in the interpretation of the overall results (Creswell, 2014).

4.2 Research Questions

The dissertation aims to answer the following research questions (RQ):

Central RQ:

Is the usage of a computer-based mode of achievement tests justified in the context of Czech tertiary education of the first year English language learners?

The central research question can be further divided into quantitative, qualitative and what Tashakkori and Creswell (2007) call a 'hybrid' or 'integrated' sub questions (p.208).

Sub-RQs:

1. Are there significant differences in the scores from the Computer-based test (CBT) and Pencil and paper-based test (PPT) modes?
2. Do the scores from the CBT and PPT modes differ in terms of gender?
3. Do the scores from the CBT and PPT modes differ with respect to the question type?

4. What are the advantages and disadvantages of the two modes of administration as viewed by students?
5. What are the students' preferences concerning the two modes of administration?
6. Is there any clear link between the student preferences and the scores they gain?

The first three sub-questions concern the quantitative data collected, sub-questions 4 and 5 are associated with the qualitative data and the last content-focused mixed methods research sub-question combines the quantitative and qualitative strands during the interpretation.

4.3 Participants

All the participants involved in the study were first year full time students studying English at the Department of English and American studies at the Faculty of Arts, Masaryk University in the Czech Republic. Since the achievement tests under scrutiny are compulsory for all first year students studying at the department, the whole population was involved and no sampling procedures were necessary. In accordance with Soukup and Kočvarová (2016), who claim that with numbers up to a hundred, the whole population rather than a sample should be worked with, and given that during each phase of the research, the number of participants was around a hundred, never higher than 150, the decision was made to work with the whole population of the first year. Self-report feedback forms administered immediately after the tests were also filled in by all research participants. This enabled the researcher to get a complete picture of the area researched for the given population. The participants' ages ranged from 19 to 25 with the majority being in their early 20s. A total of 378 undergraduate non-native speakers of English took part in the longitudinal research carried out over the period of 3 years, i.e. 2014-2016. Most of the students were Czech and Slovak and during each phase, there were also approximately 5%-8% of international students, studying at the department as Erasmus/ visiting students. The nationality variable is not further observed in this study as the population is not stratified and drawing conclusions about a few

individuals would lack validity. The researcher worked with first year students only, so there were different participants each year. See Table 8 for the breakdown of participants at each phase.

Phase/ Year	Total	Female	Male
Pilot/ 2014	138	88	50
Study 1/ 2015	114	71	43
Study 2/ 2016	126	86	40

Table 8: Participant Numbers

All the participants in all the phases signed an active consent form allowing the researcher to work with all the data obtained. In the Pilot and Study 1, there was a certain drop out of students as there was a 6-week gap between the first and second test administration and thus some students only participated in half of the research (due to illness, unspecified absence or termination of their studies) and have thus been eliminated from the quantitative part of the study. This concerns 8 students in the Pilot stage (i.e. the original number of students was 146) and 15 students in Study 1 (i.e. originally 129).

4.4 Research Apparatus

The quantitative research instrument, i.e. the computer-based tests (CBTs), originally derived from the pencil and paper-based (PP) ones, have been used at the Department since 2008. The question types correspond to the types discussed in the theoretical part of the dissertation (see 3.2.9 for more details) and include multiple choice and short answer questions. Their advantages as well as potential pitfalls have also been covered in 3.2.9. All the individual items have been carefully analysed over the years, modified, non-functioning distractors and questions eliminated, new items added, tested and calibrated, which resulted in a large item bank of well-functioning items. The tests can thus be considered a valid, reliable and standardized measure of the domains tested. The tests are used to check the

students' learning and are based on course book material covered in class or at home by self-study. The domains tested include specific areas of grammar and vocabulary and the students know exactly what to prepare for since the content is clearly specified (For more details on the test type see 2.3 in the theoretical part and for detailed test specifications see Appendix 0). The researcher is aware that the achievement tests do not focus on all the skills and systems and thus not all aspects of communicative competence are examined. However, it is not the aim of this dissertation to assess the students' overall level. The focus is on the mode of delivery of the achievement tests and its potential differences.

The computer-based achievement tests can also be described using Suvorov and Hegelheimer's scheme presented in the theoretical part (see 3.2 for details):

#	ATTRIBUTE	CATEGORY
1	Directionality	Linear (and Semi-adaptive in Pilot)
2	Delivery format	Web-based
3	Media density	Single medium
4	Target skill	Discrete-point
5	Scoring mechanism	Human-based and exact answer matching
6	Stakes	Medium-stakes
7	Purpose	Curriculum-related
8	Response type	Selected response and constructed response
9	Task type	Selective and productive

Table 9: CBT Description

(Taken and adapted from Suvorov and Hegelheimer, 2014, p.2)

The qualitative research instrument used in the study is an introspective method in the form of a self-report feedback form (Nunan, 1992), which was administered to all students immediately after the test to find out how they felt about the test mode intervention. Its format is simple and asks the students two to four open-ended questions (depending on the stage of the research) about the mode

of the test they have taken, a closed multiple-choice question to state their preference regarding the mode and a space to provide other comments. The qualitative analysis is approached from the realistic as opposed to the narrative perspective (Silverman, 2006). As recommended by Švaříček et al. (2007), a close link between the statements made and the original source of data is strictly observed, hence, for example, the students are quoted directly without linguistic alterations of their answers (i.e. no rephrasing, no mistakes corrected, no expressions translated, etc.). Given the students' high level of English and the fact that the whole degree programme is taught in English, the feedback forms were collected in English. In accordance with Švaříček et al. (2007), who warn against the mere pointing out of interesting data collected, the researcher has decided to subject the data to a systematic analysis and interpretation. The analysis consisted of three phases of coding, namely open, which Dornyei (2007) considers to be 'the first level of conceptual analysis of the data', axial, which aims to make links between categories, and selective, which pinpoints the core categories to be focused on (pp.260, 261). Through this instrument, the most valuable data and insights were assembled into the area under scrutiny.

Furthermore, a questionnaire asking the students' opinions re the two testing modes has been devised based on the qualitative data gained. It has been administered to students taking part in Study 1 and Study 2 in order to find out whether its results agree with the qualitative data provided by the same students. In the present dissertation, it is not to be viewed as a research instrument of the same importance as the two mentioned above, but rather as a work in progress for future studies. The researcher's aim is to develop, pilot and validate this instrument for future studies dealing with different modes of test administration. A more detailed description of the questionnaire will be provided in 5.7.

4.5 Research and Data Collection Procedures

The research process consists of three stages, namely the Pilot, Study 1 and Study 2. In all the stages, all the participants (who were different each year) took both the computer-based and the pencil and paper-based versions of the achievement test. Apart from the tests, participants were also asked to fill in self-report feedback forms on paper immediately following each test. In Study 1 and 2, the students also filled in an online questionnaire in order to aid the verification and validation of this research instrument for future studies conducted.

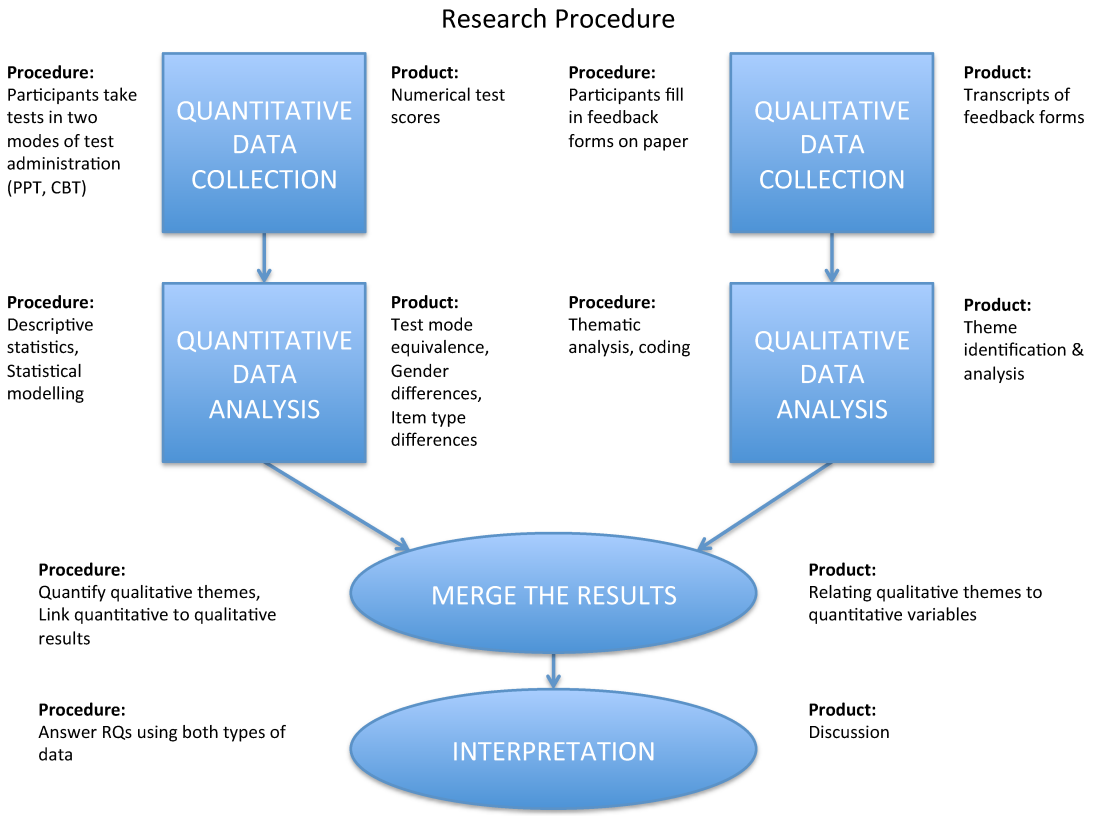

Figure 2: Research Procedure

Details specifying the data collection procedure:

PILOT: 2014

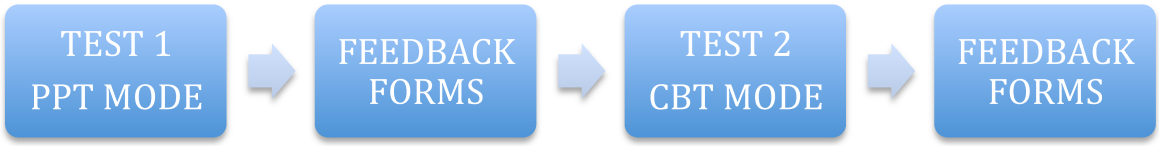

Figure 3: Pilot - Data Collection Procedure

- All students take pencil and paper-based achievement Test 1 and fill in open-ended feedback forms related to the PPT mode of administration
- Gap of 6 weeks
- All students take computer-based achievement Test 2 and fill in open-ended feedback forms related to the CBT mode of administration

STUDY 1: 2015

Figure 4: Study 1 - Data Collection Procedure

- Students are randomly divided into two halves
- The counter-balancing technique is employed, i.e. one half takes the pencil and paper-based version of Achievement Test 1, the other half takes an identical version of Achievement Test 1 but in the computer-based mode. All students fill in feedback forms commenting on the mode that they took the test in.
- Gap of 6 weeks
- The two halves swap, i.e. the half which took the pencil and paper-based test first now takes the computer-based Achievement Test 2 and vice versa. All students fill in feedback forms commenting on the mode that they took the test in.
- Gap of 4 weeks

- Students fill in an online questionnaire to confirm their qualitative data collected and also help the researcher pilot the instrument

STUDY 2: 2016

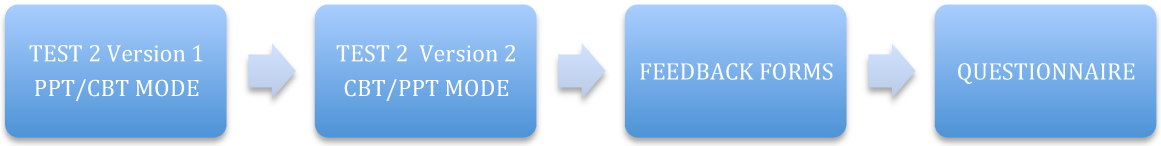

Figure 5: Study 2 - Data Collection Procedure

- Students are randomly divided into two halves
- No gap of 6 weeks, students take two versions of one test in both modes on one day
- Half the students takes the PPT mode of Version 1 of Achievement Test 2, half takes the CBT mode of Version 1 of Achievement Test 2
- Students swap
- Half the students takes the CBT mode of Version 2 of Achievement Test 2, half takes the PPT mode of Version 2 of Achievement Test 2
- All students fill in feedback forms about both modes of test administration
- Gap of 4 weeks
- Online questionnaire (same as in Study 1)

4.6 Limitations

- The design of the research was slightly modified throughout as the researcher always tried to improve on what has not worked in the previous stage to get more valid and reliable data. While this is encouraged in qualitative research design, it can be looked down on in quantitative research design. However, since this was a longitudinal mixed methods research with three independent stages each involving different participants,

it should not be viewed as a flaw but rather as a way of gaining a better insight.

- Richness of qualitative data has been sacrificed to some extent at the expense of keeping the whole population rather than sampling it for the qualitative strand – hence the usage of the feedback forms for all students instead of, for example, interviews with a few individuals, such as extreme cases, etc.
- The 6-week gap in the Pilot and Study 1 was there not to interfere with the usual set up of the course but to portray the realistic situation of students taking Test 1 halfway through the semester and Test 2 at the end. However, this proved rather problematic as although the emphasis was put on comparing the two modes of test administration and the research design in Study 1 was slightly modified to fit the purposes better, extraneous variables, such as change in ability over time, different amount of time spent studying, and most importantly different content area rendered some of the data rather questionable. That is one of the main reasons for conducting Study 2 in 2016 and administering two versions of the same achievement test in the two modes to all students during one test sitting.
- Doing research with first year full time university students only can be seen as limiting. The age of the participants probably has an effect on their attitudes to the CBT mode. As discussed in the theoretical chapter (3.1), digital residents most likely perceive the computer mode differently from digital visitors. This could definitely be investigated further.
- The questionnaires were not filled in by all participants and originally done anonymously, so it was not possible to match them to the individual scores and the qualitative data gained in Study 1. This changed in Study 2. However, given the inconsistency, lower numbers and limited space, the questionnaire is to be viewed as neither a finalized nor validated research instrument.
- The researcher was planning to include teachers' attitudes towards using the CBT mode as they are likely to influence how the students feel and repeatedly collected their insights in a focus group, however given the scope

of the dissertation, it was decided that this aspect be included in a subsequent study.

5 DATA ANALYSIS

This chapter presents and analyses the quantitative and qualitative data collected. In terms of the quantitative data, descriptive statistics, reliability estimates and statistical modelling represent the core of the analysis. Regarding the qualitative data, the process of thematic analysis and coding needs to be explained in more detail. In the Pilot stage, initial open coding was carried out by means of conceptual analysis of the student answers. Various themes were underlined in the students' feedback forms and then a more focused stage of coding took place. Categories were developed and given mostly descriptive codes. In Study 1 and Study 2 all the feedback forms were numbered (that is how the students are referred to in the analysis) and transcribed. The coding was carried out in Microsoft Excel and the themes were also related to one another. For example, the theme Technology-related was refined in Study 2 to include aspects of Technical difficulties, Log-in processes, Screen, Keyboard, Computer noise, etc., all linked to Technology as the theme title suggests but coded separately. (See Appendix 4A for the list of themes and codes.) In the last section of this chapter, the development of a new research tool based on the qualitative data gained is briefly discussed.

Given the amount of data worked with, it would be virtually impossible to include all the accompanying materials in the Appendices, which is one of the main reasons why a shared storage space on a drive was created, where all the test data, descriptive statistics, outputs for the statistical models, coded transcripts of the feedback forms, preliminary questionnaire analyses, etc., can be accessed. (See Appendix 0 for the link.)

5.1 Pilot Quantitative Data Analysis

In the Pilot stage, all students took the pencil and paper-based version of Achievement Test 1 first and the computer-based version of Achievement Test 2 second. Since the Pilot was carried out in the students' first semester of their first year, it was believed that students should start with a mode they were familiar with (i.e. the pencil and paper-based test – PPT) and then proceed to the mode used at the Department, i.e. the Computer-based test (CBT).

5.1.1 Descriptives, Limitations and Discussion

The descriptive data (*Table 10*) shows that the mean of the test scores is higher on the PPT version than on the CBT version, though the difference is very small, only 1.2 out of a total of 100 points. The standard deviations of the two versions are also very close, implying that the scores are spread out to a similar degree for the tests taken under two different conditions.

Mode	N	Mean	SD	Median	Min	Max	Range
Computer-based	138	62.9	13.5	64.2	25.2	90.5	65.4
Pencil & paper	138	64.1	12.4	66.0	33.0	89.0	56.0

N = number of students, SD = standard deviation

Table 10: Pilot - Test scores by mode of administration

Interestingly, the range, which describes the difference between the highest and lowest test scores, differs by 9.4 points, which may seem quite a considerable difference at first sight. However, given that there were 50 questions in each test, and each question was worth 2 points, the difference in range corresponds only to about 5 questions out of 50.

Though the overall pattern of the score distribution is similar, two outliers have been identified in the CBT version, who scored markedly lower than in the PPT. This will be further explored below. This is also visible in the difference

between the minimum scores, with the PPT lowest score being 33 points and the CBT lowest score being 25.2 points.

The paired sample t-test was conducted to compare the total test scores in the PPT and CBT modes. There was no significant difference in the scores in the PPT mode (M=64.1, SD=12.4) and CBT mode (M=62.9, SD=13.5); $t(137)=1.05$, $p = 0.29$. This suggests no significant relationship of testing mode and test score. (See Appendix 1A for more details.)

It needs to be stated here that it was not possible to calculate reliability estimates for the Pilot versions of the tests since the item analysis was not carried out for the following reasons. First, the students were given the PPTs back for feedback purposes, so that they could learn from their mistakes and the researcher thus did not have them at hand to transcribe and analyse the results. Second, for the CBT mode, the tests were randomly generated from a large item bank, which on the one hand portrayed the realistic situation and captured one of the fundamental advantages of the CBT mode (as discussed in the theoretical part, see 3.3) but on the other hand, made the reliability estimates impossible as each student had a different combination of the individual items. Testing reliability of sets of random items is more suitable in studies with very large numbers of participants, otherwise not all items occur enough times in the test to be analysed statistically.
Not being able to calculate test reliability was one of the main reasons for changing the research design in the subsequent studies.

When male and female test scores are combined, as was the case in the statistics described above, the test scores resemble a normal distribution and there are no major discrepancies. However, it is in the interest of the researcher to point to the problematic elements, not always visible at first sight and since the researcher wanted to examine possible gender differences, the following histogram (Figure 6), which shows the distribution of test scores by gender and mode of administration, is presented. (See also Appendix 1B for descriptive statistics of tests scores by mode of administration and gender.)

The horizontal axis displays the number of points gained (i.e. the test score) and the vertical axis indicates the number of students who gained them. Each vertical bar represents the amount of students who obtained the score in a given range. The plot is divided into four facets, where each facet represents one subgroup of test takers. Histograms for women (n=88) occupy the left column of the grid and men (n=50) are on the right. The scores in the PPT version of the test are displayed in the first row, the CBT in the second.

Figure 6: Pilot - Distribution of test scores

Nobody scored below 20 points, which is to be expected given it is an achievement test. The most notable feature of the faceted plot above is the difference between test scores obtained by women compared to men. This difference is not in the overall height of the bars – since, as was mentioned, this merely represents the amount of test-takers of a given gender, and there were fewer men amongst the test-takers - but in the *shape* of the distribution. While the expected shape would resemble a bell curve with highest bars around the mean, and indeed it does so in the case of women, the male scores are notably different. Neither of the tests (PPT or CBT) produced a bell-shaped distribution with a clear

central tendency, which renders descriptive statistics, such as mean, very poor representations of its characteristics. This means that a direct comparison of female and male scores is problematic and not very meaningful, and makes further statistical testing difficult.

As for women's scores, the shape is much closer to an ideal Gaussian distribution with a negligible left skew. The only irregularity worth mentioning is the solitary bar on the left (bottom left facet). These outliers represent a minor deviation from normality, and are explained further in the text.

The possible explanations for the male plateau distribution of test scores remain uncertain. One could hypothesize that the number of male students was considerably lower, which often leads to non-normal distributions. Faulty measurement has also been considered but since it was not the case with the female scores, it seems rather unlikely. Furthermore, all the scores have been carefully checked. Cheating during the test as a potential explanation has been eliminated due to the fact that when taking the CBT (in which each student had a different version of the test drawn from the large item bank), the students would not be able to cheat the same way as in the PPT (where all students had the same version), which would result in the group of cheaters having high PPT scores and low CBT scores. This would all be visible on the scatter plot (see Appendix 1C), where the values would likely form a recognizable cluster of high PPT but low CBT values. However, this is not the case.

The main problem with the pilot data is that the male and female score distributions differ so dramatically. This makes the comparison between male and female scores problematic and was one of the reasons why the design of the research was changed in Study 1.

One last aspect worth mentioning is the case of the outliers clearly visible in the bottom left corner of the histogram (Figure 6). The two females scoring considerably lower in the CBT mode have already been mentioned above. A closer look at these individuals reveals that they are visiting students whose grammatical and lexical competences are lower than those of our students' and while they still

somehow coped with the first achievement test, the second achievement most likely proved too difficult for them. This would suggest that the difference is not caused by the mode of administration but rather by the content of the tests. Another feasible explanation would be that the two female students cheated in the PPT and thus got a higher score, while they were not able to cheat in the CBT and thus gained a considerably lower score as mentioned above (see Appendix 1C).

Overall, regarding the answers to the research questions:

1. The test scores do not significantly differ across the two versions.

2. & 3. For the reasons stated above, it is not possible to compare male and female student results or potential differences in the question types.

All these concerns are addressed in the following studies, namely Study 1 and Study 2.

5.2 Study 1 Quantitative Data Analysis

In Study 1, the design of the research was altered and counter-balancing order was employed in order to eliminate sequencing effects. These can happen when the two tests are taken at different times, which makes it impossible to know whether the observed difference in scores is due to the testing mode or the simple fact that one version of the test contained more difficult items than the other. Counter-balanced order ensures that whatever the effect of difficulty, it will get evenly distributed between the two testing modes. Therefore, half the students took the pencil and paper-based version of Achievement Test 1 first (referred to as Version 1 below) and the computer-based version of Achievement Test 2 second (referred to as Version 2 below) and the other half vice versa, i.e. the CBT version of Achievement Test 1 first (referred to as Version 1 below) and the PPT of Achievement Test 2 second (referred to as Version 2 below).

In contrast to the Pilot, in Study 1 students did not take the same test in the same mode. Furthermore, while in the Pilot, only the total scores were worked with as the item data was not available, in Study 1, there were 38 questions and each question received a mark from 0-1. All multiple choice items were graded dichotomously, i.e. 1 for correct answer and 0 for incorrect answer, while short

answer items also allowed values in between (0.5 for instance for partially correct answers). Having scores for individual items, the reliability of the test can be assessed, which was not possible in the Pilot because of the lack of item data. The present tests have been analysed with respect to the two distinct subtests, namely Destination (measuring aspects of grammatical competence) and Vocabulary (measuring aspects of lexical competence). More data is therefore gained and given that there are two versions, two subtests and two groups, the descriptive statistics described below is more detailed than in the Pilot stage.

5.2.1 Reliability Estimates

When measuring people's abilities/knowledge, the validity and reliability of the tools needs to be assessed. While content validity can be determined by domain experts, reliability has to be estimated given the present data. In the present sample, the Cronbach's alpha coefficient (Cronbach, 1951) has been used, which is appropriate for tests with a unidimensional structure (i.e. discrete-point tests as discussed in 2.3.2). Since the contents of the test were divided into two narrowly defined groups (Destination and Vocabulary), it was assumed that each of the subtests meets requirements for a unidimensional structure. Furthermore, a multidimensional item structure, if estimated by tools that expect a unidimensional structure, would result in a lower reliability estimate so the need for advanced methods like factor analysis or structural modelling would be apparent.

However, this was not the case. The estimates of reliability in all subtests regardless of mode fell within an acceptable range (see Tables 11 and 12 below). A threshold value of .7 is sometimes cited in the literature (see, for example, Tavakol & Dennick, 2011), but small deviations from this should not be seen as critical. A conservative point of view would point out Version 2 of the subtest "Vocabulary" where alpha = .6 only. This could also be related to the smallest inter-item correlation (i.e. how similarly the items do on a test) – in tests with lower inter-item correlation, there may be some items that function differently from others (i.e. they measure a different dimension of ability or are simply not good at measuring what they are supposed to), lowering unidimensional reliability estimates like Cronbach

alpha. But the value is not seen as critical enough so as to warrant further investigation at this point.

There are two estimates of reliability, Cronbach's alpha and Guttman's lambda6, both measures of unidimensional internal consistency. The inter-item correlations (r) indicate how similar were the scores between individual items. For scales with good reliability, lower inter-item correlations suggest less redundant items.

Subtest	Testing mode	alpha	G6	Inter-item r	Mean	SD
Destination	Computer-based	0.62	0.71	0.08	0.65	0.15
	Pencil & paper	0.72	0.80	0.13	0.68	0.17
Vocabulary	Computer-based	0.74	0.83	0.13	0.66	0.16
	Pencil & paper	0.69	0.81	0.10	0.72	0.14

Table 11: Study 1 - Reliability estimates for Version 1 subtests

Subtest	Testing mode	alpha	G6	Inter-item r	Mean	SD
Destination	Computer-based	0.72	0.80	0.13	0.77	0.16
	Pencil & paper	0.76	0.84	0.15	0.76	0.17
Vocabulary	Computer-based	0.67	0.81	0.11	0.76	0.14
	Pencil & paper	0.60	0.72	0.09	0.74	0.14

Table 12: Study 1 - Reliability estimates for Version 2 subtests

5.2.2 Descriptives

Table 13 below indicates that the mean of the test scores is slightly higher in the second version, i.e. Achievement Test 2. The magnitude of standard deviations

does not suggest substantial deviation from normality. It seems that the Destination subtest of Version 1 and the Vocabulary subtest of Version 2 in particular produced the most diverse results, the former being more challenging for the students than its Version 1 counterpart and the latter being the easiest of all subtests. This is further supported by that subtest having the highest median of 16 and the smallest range of 12, with the minimum score being 8 points.

Version	Subtest	N	Mean	SD	Median	Min	Max	Range	Skew
Version 1	Destination	114	11.9	2.9	12.5	3.5	18.0	14.5	-0.6
	Vocabulary	114	13.9	3.1	14.0	1.0	19.8	18.8	-1.1
Version 2	Destination	114	14.0	2.8	14.5	5.0	18.0	13.0	-0.9
	Vocabulary	114	15.9	2.3	16.0	8.0	20.0	12.0	-0.9

N = number of students, SD = standard deviation
Table 13: Study 1 - Test scores by version

When looking at Table 14 below, which shows the overall test scores by group, i.e. the group which took the CBT first as opposed to the group which took the PPT first, one can notice that the latter overall outperformed the former, though by rather a negligible margin. Standard deviations are very close again and except for the very low minimum score gained in the Vocabulary subtest of the Computer first group (to be elaborated on later in greater detail below), the medians and ranges are all very similar, suggesting no large differences between groups or deviations from normal distribution.

Group	Subtest	N	Mean	SD	Median	Min	Max	Range	Skew
Computer first	Destination	116	12.8	3.1	13.0	5.0	18	13.0	-0.5

Computer first	Vocabulary	116	14.6	3.1	15.0	1.0	20	19.0	-1.6
Pencil & paper first	Destination	112	13.1	3.0	13.6	3.5	18	14.5	-0.7
Pencil & paper first	Vocabulary	112	15.2	2.7	15.5	7.0	20	13.0	-0.5

N = number of students (each student took 2 versions, i.e. 58x2=116 students and 56x2=112 students), SD = standard deviation
Table 14: Study 1 - Test scores by group

For descriptive statistics of test scores sorted by gender, which were very similar and even showed identical medians, and group and mode, see Appendices 2A and 2B respectively. Overall, the PPT mode across the two versions has slightly higher mean than the CBT mode.

The histogram below (Figure 7) displays the distribution of test scores by gender and mode. Once again, the horizontal axis shows the number of points achieved (i.e. total score) and the vertical axis the number of students who gained them. Therefore, the lower bins in the male facet do not mean lower results, just fewer males. If the two groups of scores were different, it would be indicated by a shift of the whole distribution either left or right.

Figure 7: Study 1 - Test Scores by gender and mode

Contrary to the Pilot, there are no pronounced differences or discrepancies in terms of gender score distribution. Furthermore, since this is a criterion-referenced achievement test, the majority of the data points is expected to be located to the right of the cutoff point (passing grade of 12). Since the cutoff is not in the middle of the score range but further to the right, there should be less variability on the right side of the histogram than on the left (as more scores are concentrated in a smaller range). This is why on the left side, the numbers of people achieving lower-then-cutoff scores are more spread out, resulting in lower bins and a light tail on the left side of the histogram. Apart from this minor deviation, the score distribution resembles normal distribution with a clear central tendency around the mean of the scores and does not differ greatly with respect to the mode (CBT/PPT). For a scatter plot displaying a correlation between the two modes of test administration, see Appendix 2C.

5.2.3 Comparing Mean Test Scores

Here the statistical significance of mean score differences of interest, i.e. the effect of testing mode and gender on test scores is assessed. To exemplify the differences we are interested in, Figure 8 shows mean scores in different subgroups.

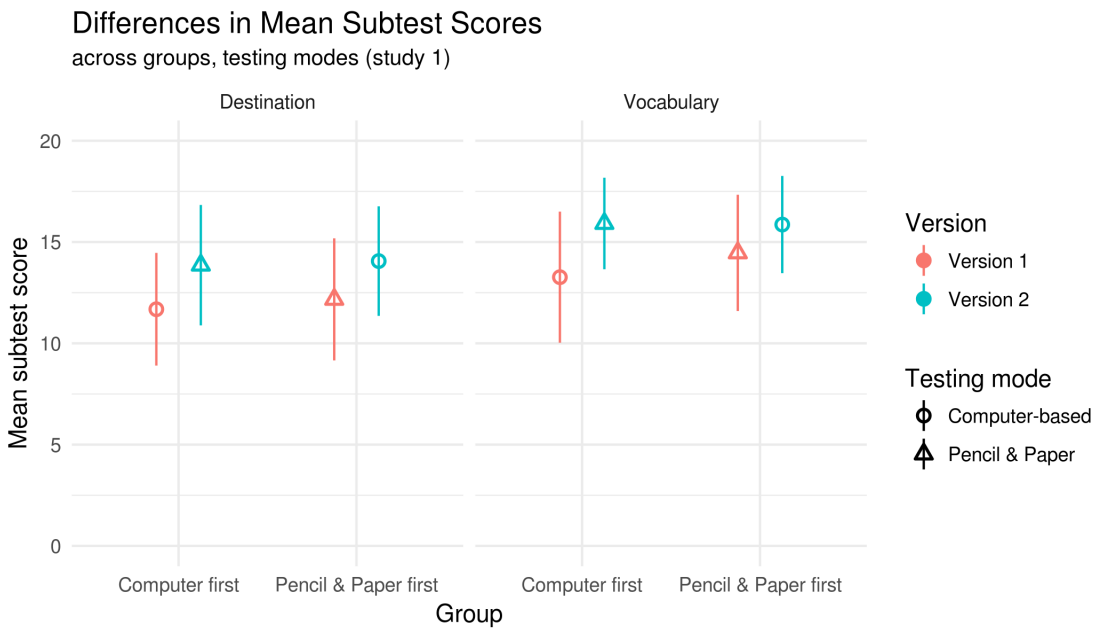

Figure 8: Study 1 – Subtest Means

Each facet on the plot above shows mean scores for the given subtest (the left facet displays scores for Destination subtest, the right one for Vocabulary. Circles and triangles represent CBT and PPT versions of the test, respectively. As mentioned above, the students took two tests in the course of the study: *Version 1* (red) and *Version 2* (blue). The order in which they took these tests was counterbalanced as follows:

- The first half of the students started with CBT (so their mean scores are indicated by red circles) and continued with PPT (blue triangles)
- The second half of the students did the exact opposite, i.e. they started with PPT (red triangles) and continued with CBT (blue circles)
- Each version was therefore carried out in two modalities (CBT vs PPT)
- The items were the same in both modalities

- However, they were different in Version 1 with respect to Version 2
- The number of items was identical in both versions

First, it is important to note that mean scores in *Version 2* (blue) are higher than in *Version 1*, which means that the difficulty of *Version 2* was lower. This was already mentioned in the section detailing descriptive statistics. Furthermore, the Vocabulary subtest (right facet) seems to show higher mean scores overall compared to the Destination subtest (left facet). A more subtle feature of this plot is that, while in *Version 2* (blue), CBT and PPT scores seem to be almost identical, they somewhat differ in *Version 1* (red). The PPT version of the tests, represented by triangles, yielded higher mean scores (more so in the Vocabulary subtest).

This would certainly be interesting if proven statistically significant. However, the fact that this difference seems to show only in one version of the test and not the other suggests that the two versions are not equivalent in some way (the possible reasons for this, such as different counts of short answer versus multiple-choice questions, or human versus computer marking, have been presented above), and it is thus advisable to approach the results obtained by statistical methods with caution.

Statistical model 1: The effect of group (CBT first vs PPT first)

In order to test which variables influence the mean subtest score, the following variables have been factored in:

- **testing mode:** this is the primary interest, as it is believed that it may affect test scores
- **group:** it should be confirmed that those who started with the CBT mode and followed with the PPT do not differ in terms of scores from those who started with the PPT and followed with the CBT (having accounted for differences in testing mode and version)

- **test-taker:** since each student has effectively done two tests (PPT, CBT), the data that comes from these tests is not independent. To account for interpersonal variability in ability/knowledge, test-taker IDs are included in the model, which thus accounts for student's individual test scores
- **version:** due to different item structure, the version of test can affect scores in different ways depending on a particular combination of the other variables.

These are the predictors in the model with **subtest score** as the outcome. A mixed-effect model with fixed effects of *testing mode (CBT vs PPT)* and *group (CBT first vs PPT first)* and random effects of *version (1 and 2)* and *test-taker (student id)* was constructed.

This method is implemented in R by Bates et al. (2015). Since the researcher is also interested in statistical significance, a p value for fixed effects is included in this model. Obtaining p values in such models is not a straightforward procedure and is sometimes discouraged, but for a repeated measures design, Satterthwaite approximation is often used (Kuznetsova et al., 2016). For a complete output of the modelling procedure, see the link to the drive in Appendix 0. This procedure is also used for all subsequent statistical models in 5.2 and 5.3.

Results: Destination subtest

- The resulting model yielded an intercept of 12.7 (SE = 0.8) - this represents the average score *without* the influence of testing mode or group.
- The effect of **group** was not significant (beta = .34, SE = .49, p = .48) which suggests that belonging to a certain group did not significantly affect test scores.
- The effect of **testing mode** was also not significant (beta = .14, SE = .21, p = .5). This means that the test scores on the Destination subtest were not affected by testing mode.

Results: Vocabulary subtest

- The resulting model yielded an intercept of 14.27 (SE = 0.8) - this represents the average score *without* the influence of testing mode or group.

- The effect of **group** was not significant (beta = .57, SE = .43, p = .19) which suggests that belonging to a certain group did not significantly affect test scores.
- The effect of **testing mode** was significant on a less conservative level of alpha < 0.05 (beta = .57, SE = .43, p = .016). This translates into improving test scores by .57 if taken in the PPT mode - note however that this would not be considered significant on a more conservative level (alpha < .01).

While the small size of the effect of testing mode is to be expected (it is not expected that a computer test would be *radically* different in terms of difficulty), the *practical* significance of a result that does not even amount to a single item should be questioned. Furthermore, the statistical significance is only reached if liberal threshold is chosen (alpha < .05), which is not very convincing.

Statistical model 2: The effect of gender

The following model seeks to find out whether there are differences between the mean subtest scores with respect to gender. The model has the same design as Statistical model 1, with the exception of the second fixed effect, which in this case is gender (male vs female).

Results: Destination subtest (mean subtest scores)

- The resulting model yielded an intercept of 13 (SE = 0.8) - this represents the average score *without* the influence of testing mode or gender.
- The effect of **testing mode** was not significant (beta = -0.08, SE = 0.27, p = 0.75). This would suggest that students do not perform differently in the PPT/CBT mode.
- The effect of **gender** was not significant (beta = -0.41, SE = 0.55, p = 0.46) which suggests that men and women do not differ in test scores significantly.
- The interaction between **testing mode** and **gender** was also not significant (beta = 0.6, SE = 0.45, p = 0.17). This suggests that there is no difference in how men and women are affected by testing mode.

Results: Vocabulary subtest (mean subtest scores)

- The resulting model yielded an intercept of 14.4 (SE = 0.8) - this represents the average score *without* the influence of testing mode or gender.
- The effect of **testing mode** was marginally significant (beta = 0.7, SE = 0.32, p = 0.03). This would suggest a very slight effect of testing mode (0.7 point higher mean score in the PPT mode), but only on a less conservative alpha < 0.05.
- The effect of **gender** was not significant (beta = 0.33, SE = 0.52, p = 0.52) which suggests that men and women do not significantly differ in test scores.
- The interaction between **testing mode** and **gender** was also not significant (beta = -0.21, SE = 0.53, p = 0.69). This suggests no difference in how men and women are affected by testing mode.

Similarly to Statistical model 1, the only marginally significant difference would be in the Vocabulary subtest with respect to the testing mode. Neither gender differences in test scores in general, nor differences related to the testing mode and gender have been discovered.

5.2.4 Study 1 Discussion, Limitations and Conclusions

All in all, both the descriptive statistics and the statistical models showed only minor differences in student performance. These could be explained by the fact that these particular students were more familiar with vocabulary topics covered in Achievement Test 2 (i.e. Version 2) or just in general studied harder for the 2nd test, which can be further supported by the Destination subtest displaying better scores too, although as mentioned above with no statistical significance.

The fact that the PPT scores were slightly better than the CBT ones, especially in the Vocabulary subtest, could be accounted for by the Vocabulary subtests containing more short answer questions. Human raters could have been more lenient in marking those than the computer, even though the answer key was identical and the marking procedures standardized. For example, with unclear spelling when the answer was handwritten, some points could have been awarded in the PPT version. In the case of the computer-based scoring of short answer questions (also checked by human raters), there was no space for doubt when the

answers were mistyped. Subsequently, since the Destination subtests included more multiple-choice questions, the scoring was more straightforward and less prone to misinterpretations.

A few outliers have been identified, well visible in both the histogram and scatter plot as well as pointed out in the descriptive statistics above. Similarly to the Pilot, these are mainly visiting students studying at the Department for one semester. Qualitative data provided by these outliers might shed some more light on the potential reasons for their poor performance in either or both of the modes and/or versions.

Last but not least, the unequal proportion of the open-ended short answer versus multiple-choice questions across the two versions (i.e. Achievement Test 1 and 2) needs to be mentioned as a limitation of Study 1. Although the two modes of test administration (PPT and CBT) of the same version contained always the same items, the two versions themselves were not structurally identical, as Version 2 included more multiple-choice questions. One could thus hypothesize that students perform differently in open-ended questions as opposed to multiple-choice in either of the modes, which can influence the results of the individual subtests. This will be further investigated in Study 2.

Structural differences in versions pose a threat to inter-method reliability (i.e. whether the two versions of a test give the same result when assessing ability/knowledge of the same individual). One of the main reasons for conducting Study 2 was making sure that the number of the same type of questions is identical across the two versions of the tests, not only across the two modes of administration. A more detailed item-matching between the versions based on domain expertise related to content will also be performed. Eventually, a more detailed analysis and comparison across the individual versions will be enabled.

Overall, regarding the answers to the research questions:

1. The test scores slightly differ but only in one subtest (Vocabulary) if a less conservative alpha level is observed, but the difference does not even amount to 1 extra correct answer in the PPT mode.
2. Statistically significant gender differences with respect to the mode of test administration have not been observed.
3. Although the item data is available, for the reasons stated above, it may not be reliable to investigate potential differences in performance related to item-type means. Furthermore, content-matching for items between versions was not carried out. The gap of 6 weeks between test interventions poses another concern, since the amount of student preparation for the test may have varied over time.

The results of Study 1 did not bring very strong evidence in favour of the effect of testing mode or gender, but given the limitations, this could be attributed to the concerns raised above. These limitations will be addressed in Study 2 so as to ensure that the evidence (or lack thereof) is not caused by methodological issues.

5.3 Study 2 Quantitative Data Analysis

For Study 2, the counter-balanced design was maintained, the students were again randomly divided into two halves, however, there were certain alterations compared to Study 1. First of all, the time frame changed, i.e. all students took both the versions in both the modes on one day, instead of having 6 weeks in between. Furthermore, content-wise instead of Achievement Tests 1 and 2, two versions of Achievement Test 2 were administered, which was done to make sure that students were tested from the same material in both the modes. Finally, the two versions were identical in terms of format, each including exactly the same number of short answer and multiple-choice questions. Particular attention was also paid to standardization of human versus computer scoring. It was believed that this procedure would guarantee more comparable data.

Both versions contained two subtests again, namely Destination and Vocabulary. There were 25 questions in each subtest and same as in Study 1 each question received a mark from 0-1.

5.3.1 Reliability Estimates

The merits of the change in the design are shown already in the calculations of the reliability estimates. The reasons for and ways of calculating reliability estimates have been thoroughly described in Study 1 (see 5.2.1). Tables 15 and 16 below summarize the results.

Internal consistency

Subtest	Testing mode	Alpha	G6	Inter-item r	Mean	SD
Destination	Computer-based	0.74	0.84	0.10	0.66	0.16
	Pencil & paper	0.86	0.91	0.19	0.63	0.21
Vocabulary	Computer-based	0.73	0.85	0.10	0.63	0.16
	Pencil & paper	0.87	0.92	0.22	0.58	0.22

Table 15: Study 2 - Reliability estimates for version 1 subtests

Subtest	Testing mode	Alpha	G6	Inter-item r	Mean	SD
Destination	Computer-based	0.84	0.90	0.17	0.64	0.20
	Pencil & paper	0.73	0.84	0.10	0.71	0.15
Vocabulary	Computer-based	0.84	0.89	0.18	0.55	0.21
	Pencil & paper	0.81	0.87	0.15	0.64	0.19

Table 16: Study 2 - Reliability estimates for version 2 subtests

As the tables show, all the subtests in both the versions and modes exceed the 0.7 reliability estimates, which demonstrates good test reliability in terms of internal consistency.

Inter-method reliability

Inter-method reliability, also known as parallel forms reliability as discussed in the theoretical part (2.2.2) should demonstrate that test scores correlate in both modes and versions. Ideally, mean item scores should also correlate between the two versions, however, there is always a looser relationship among items due to the fact that their variability is greater than that of the test scores.

To estimate the inter-method reliability of both versions, a Pearson product moment correlation was used. In both subtests, the correlations were strong and significant on alpha < .01, i.e. $r_{(124)} = 0.82$, $p < .01$ for "Destination" and $r_{(124)} = 0.79$, $p < .01$ for Vocabulary.

Figure 9 provides scatterplots for whole-test scores as well as individual items. On the left, a strong relationship is apparent between the two versions of the test, which does not seem to be influenced by group (i.e. students who took the test in the CBT mode first do not differ from those who took it second). On the right, items are compared to their content-matched counterparts from the other version in terms of mean scores. The correlation is not so strong there but it is present nonetheless, and given that the variability of individual item scores must be greater than variability of whole-test scores, this is to be expected.

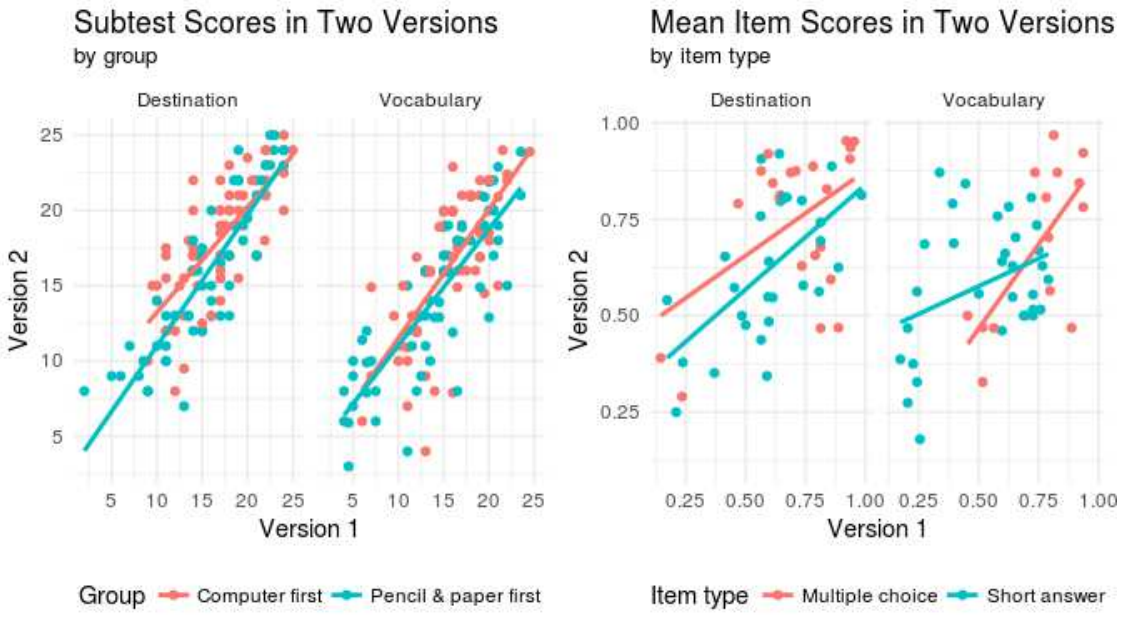

Figure 9: Study 2 – Subtest Scores and Mean Item Scores Correlations

5.3.2 Descriptives

Table 17 below demonstrates that the mean scores as well as the standard deviations are very similar across the versions. It seems that in contrast to Study 1, students performed slightly better in the Destination subtest than in the Vocabulary one in both versions. The range points to the wide variety of scores gained, including some very low and some of the highest marks possible. The median of the two Destination subtests is identical and the median of the two Vocabulary subtests differs by 0.8 points only, which is less than one correct answer.

Version	Subtest	N	Mean	SD	Median	Min	Max	Range	Skew
Version 1	Destination	126	16.2	4.6	17.0	2	25.0	23.0	-0.3
	Vocabulary	126	14.8	4.9	15.2	4	24.5	20.5	-0.4
Version 2	Destination	126	16.9	4.6	17.0	7	25.0	18.0	-0.1
	Vocabulary	126	15.2	5.0	16.0	3	24.0	21.0	-0.4

N = number of students, SD = standard deviation

Table 17: Study 2 - Test scores by version

When looking at Tables 18 and 19 below displaying the overall test scores by group and mode according to the subtests, it is apparent that the group, which started with the CBT mode, outperformed the group, which started with the PPT mode. The difference is not major but is consistent across the subtests. Slightly higher medians are also traceable in the PPT mode as opposed to the CBT.

Group	Mode	N	Mean	SD	Median	Min	Max	Range	Skew
Computer first	Computer-based	62	16.6	4.0	17.0	9	25	16	0.1
Pencil & paper first	Computer-based	64	16.0	5.1	16.0	7	25	18	0.1

Computer first	Pencil & paper	62	17.8	3.8	17.2	8	25	17	-0.3
Pencil & paper first	Pencil & paper	64	15.8	5.2	16.0	2	24	22	-0.4

Table 18: Study 2 - Destination scores by group and mode

Group	Mode	N	Mean	SD	Median	Min	Max	Range	Skew
Computer first	Computer-based	62	15.5	4.0	16.0	6	24.5	18.5	-0.2
Pencil & paper first	Computer-based	64	14.3	5.1	14.9	3	23.9	20.9	-0.2
Computer first	Pencil & paper	62	16.2	4.7	16.9	4	24.0	20.0	-0.6
Pencil & paper first	Pencil & paper	64	14.2	5.5	15.0	4	23.5	19.5	-0.3

N = number of students, SD = standard deviation

Table 19: Study 2 - Vocabulary scores by group and mode

For descriptive statistics of test scores sorted by group and gender separately, which agree with the data presented above and do not point to any major differences, especially regarding gender, see Appendices 3A and 3B respectively.

In terms of distribution of test scores, the following histogram (Figure 10) portrays balanced numbers in both modes, though the difference in the proportion of women and men (86 to 40) is apparent. Having said that, the distributions are much better in terms of normality than in the Pilot stage where the same issue was encountered, yet the male scores displayed a non-normal distribution. This is not the case here, as the shape in all four facets of the plot resembles a bell curve with a clear central tendency. One can notice a heavier tailed distribution in the female

facets and a lighter tailed distribution in the male facets, caused by the difference in the proportion of female and male students.

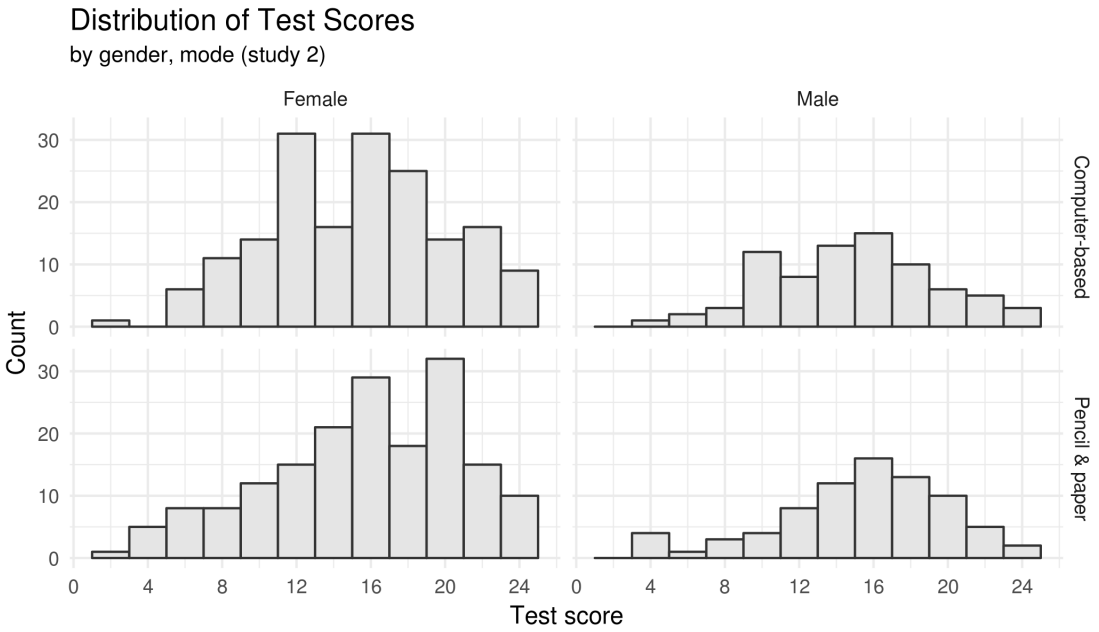

Figure 10: Study 2 - Test scores by gender and mode

For a scatter plot displaying a strong correlation between the two modes of test administration, see Appendix 3C.

5.3.3 Comparing Mean Test Scores

Here the statistical significance of mean score differences of interest, i.e. the effect of testing mode, gender and item type on test scores is assessed. To exemplify the differences we are interested in, Figure 11 shows mean scores in different subgroups.

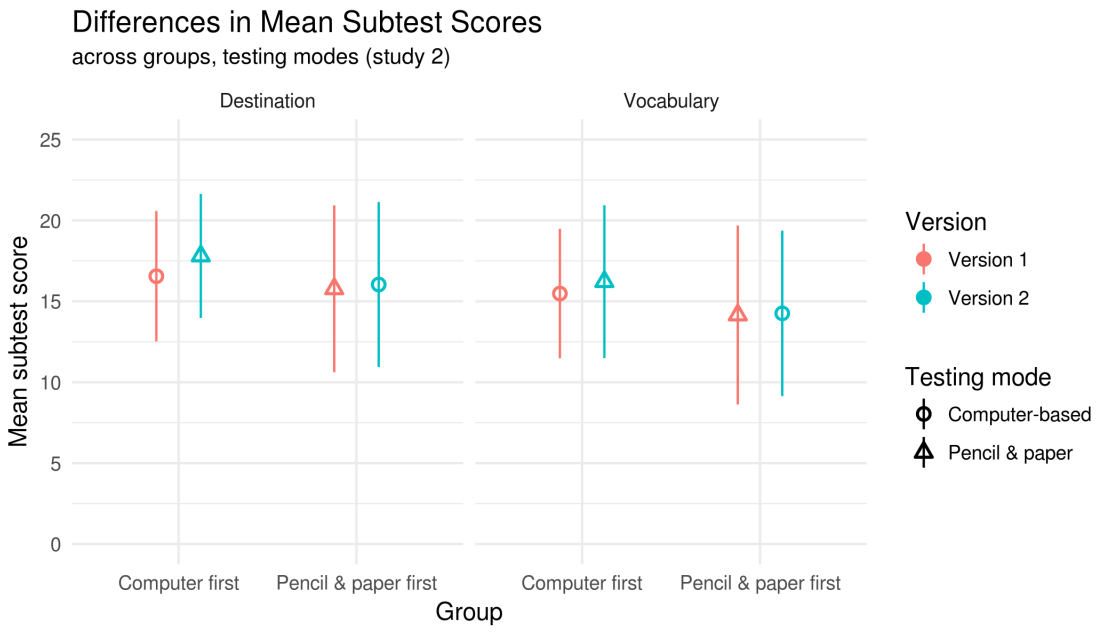

Figure 11: Study 2 - Subtest means

The logic of this plot is the same as in Study 1, where it was thoroughly described (see 5.2.3 for details). Unlike Study 1, it can be seen here that the versions are equally difficult and do not display any striking differences in test scores.

Statistical model 1: The effect of group (CBT first vs PPT first)

A statistical model to see which variables influence mean subtest score, introduced previously in Study 1 (see 5.2.3), has been used in Study 2 using the same predictors, i.e. testing mode, group, test taker, and version.

Results: Destination subtest (mean subtest scores)

- The resulting model yielded an intercept of 16.9 (SE = 0.63) - this represents the average score *without* the influence of testing mode or group.
- The effect of **group** was not significant (beta = -1.27, SE = .77, p = .1) which suggests that belonging to a certain group did not significantly affect test scores.
- The effect of **testing mode** was marginally significant (beta = .49, SE = .24, p = .04). This would suggest an increase of mean score in the Destination subtest of the PPT, however, the increase is negligible (.49 points, which roughly

corresponds to half of one item), and the p value would not suggest significance on a more conservative level alpha < .01.

Results: Vocabulary subtest (mean subtest scores)

- The resulting model yielded an intercept of 15.69 (SE = .61) - this represents the average score *without* the influence of testing mode or group.
- The effect of **group** was not significant (beta = -1.64, SE = .82, p = .26) which suggests that belonging to a certain group did not significantly affect test scores.
- The effect of **testing mode** was not significant (beta = .31, SE = .81, p = .05).

Statistical model 2: The effect of gender

Similarly to Study 1, the following model aims to find out whether there are statistically significant differences between the mean subtest scores with respect to gender. The model has the same design as Statistical model 1, with the exception of the second fixed effect, which in this case is gender.

Results: Destination subtest (mean subtest scores)

- The resulting model yielded an intercept of 16.4 (SE = 0.57) - this represents the average score *without* the influence of testing mode or gender.
- The effect of **testing mode** was not significant (beta = 0.54, SE = 0.28, p = 0.06). This would suggest no effect of testing mode.
- The effect of **gender** was not significant (beta = -0.44, SE = 0.88, p = 0.61) which suggests men and women did not significantly differ in test scores.
- The interaction between **testing mode** and **gender** was also not significant (beta = -0.16, SE = 0.52, p = 0.75). This suggests that there is no difference in how men and women are affected by testing mode.

Results: Vocabulary subtest (mean subtest scores)

- The resulting model yielded an intercept of 14.9 (SE = 0.5) - this represents the average score *without* the influence of testing mode or gender.
- The effect of **testing mode** was not significant (beta = 0.18, SE = 0.34, p = 0.57). This would suggest no effect of testing mode.

- The effect of **gender** was not significant (beta = -0.26, SE = 0.94, p = 0.77) which suggests men and women did not significantly differ in test scores.
- The interaction between **testing mode** and **gender** was also not significant (beta = 0.4, SE = 0.6, p = 0.5). This suggests that there is no difference in how men and women are affected by testing mode.

As the results above demonstrate, no significant differences have been found related to gender, nor the interaction between testing mode and gender in either of the subtests.

5.3.4 Comparing Item Means

Certain differences were noticed in terms of student performance on multiple choice versus short answer questions already in Study 1 but it would have been difficult to explore these differences any further as the versions were not identical in terms of format. Although the overall test scores do not notably differ, it is possible that the item scores do, which is why this aspect is considered. The two item types (multiple choice and short answer questions) will now be examined in more detail.

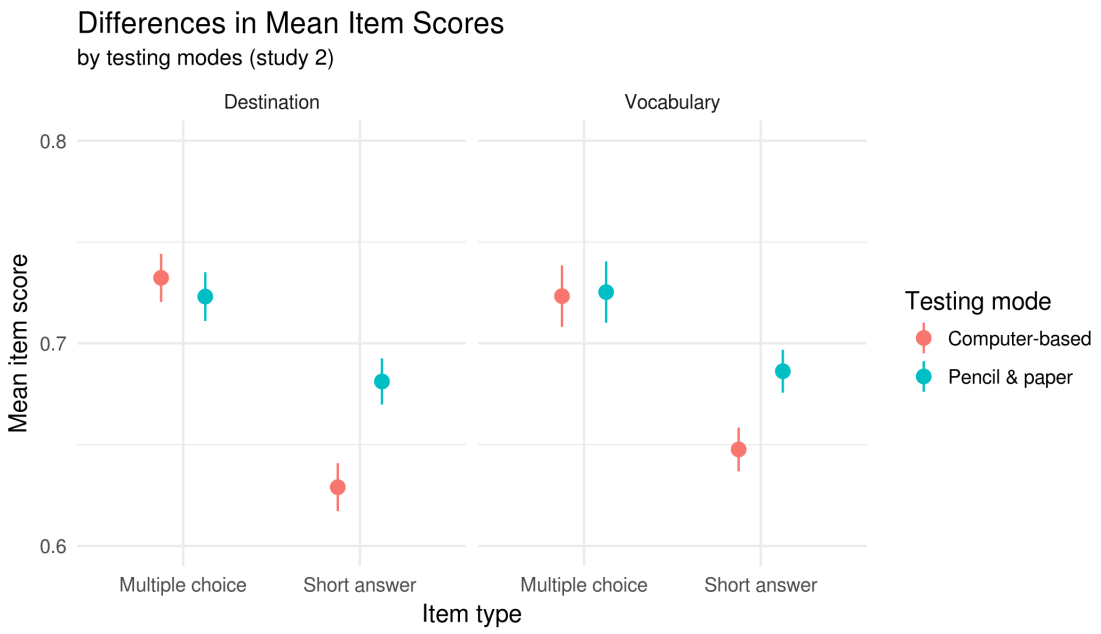

Figure 12: Study 2 - Item score means

On this plot, the facets represent the two subtests. The dots represent mean item scores for multiple choice and short answer items. They are coloured according to the testing mode - red dots for CBT, blue for PPT.

The plot demonstrates that short answer items are in general more difficult (the pairs of dots that represent short answer items are lower on the vertical axis). Furthermore, short answer items in the PPT (blue), have higher mean scores. This is true for both subtests but is more pronounced in the Destination subtest.

Statistical Model 3: Comparing item means

The significance of this difference will therefore be tested. The predictors used in Statistical model 3 are:

- **testing mode:** it is believed it may affect item scores, most likely in interaction with item type
- **item type:** it is believed it may affect item scores in one type of item depending on which testing mode was applied
- **test-taker:** since each student has effectively done two tests (PPT, CBT), the data that comes from these tests is not independent. To account for interpersonal variability in ability/knowledge, we include student ID in the model
- **version:** some variance is bound to exist between the versions, meaning that each of the two versions of the test can affect scores in a different way.

These are the predictors in our model with **mean item score** as the outcome (the value range of this variable is 0-1). A mixed-effect model with interaction between fixed effects of *testing mode (CBT vs PPT)* and *item type (SA vs MC)* and random effects of *version (1 and 2)* and *test-taker (student id)* was constructed.

Results: Destination subtest (mean item scores)

- The resulting model yielded an intercept of .73 (SE = .02) - this represents the average score *without* the influence of testing mode or item type.
- The effect of **testing mode** in itself was not statistically significant (beta = .0, SE = .01, p = .5).
- The effect of **item type** in itself was highly significant (beta = -.1, SE = .01, p < .001). This means that SA items are on average more difficult than MC items (perhaps to be expected).
- The interaction between **testing mode** and **item type** was statistically significant (beta = .06, SE = .02, p < .01). This means that when written in the PPT mode, the mean score on SA items is higher than with the CBT mode, by about 6%. Since there are 25 items on the subtest, 1 item corresponds to 4% of the test, so the effect is roughly equal to 1 and a half additional item achieved correct.

Results: Vocabulary subtest (mean item scores)

- The resulting model yielded an intercept of .72 (SE = .01) - this represents the average score *without* the influence of testing mode or item type.
- The effect of **testing mode** in itself was not statistically significant (beta = .0, SE = .01, p = .9).
- The effect of **item type** in itself was highly significant (beta = -.08, SE = .02, p < .001). This means that SA items are on average more difficult than MC items (this was observed also in the Destination subtest).
- The interaction between **testing mode** and **item type** went in the same direction as in Destination subtest, but was not statistically significant (beta = .02, SE = .02, p = .4)

On the plot (Figure 12), a difference in short answer item scores in both subtests is observed but this effect of item type is only significant in the Destination subtest. This could be due to higher variability in short answer item scores in Vocabulary, or it could have other reasons possibly explained by analysing data from the qualitative analysis.

5.3.5 Study 2 Discussion, Limitations and Conclusions

Study 2 had the most elaborate design and the results shed more light on the research area in question. The reliability estimates increased and the descriptive statistics did not point to any non-normal distributions or discrepancies in test scores. Three statistical models were incorporated examining potential effects of test group, gender and item types.

What can be seen as a limitation of Study 2 is that the research design was general. Now that the equivalency of the two modes has been demonstrated, future studies could focus on the specific aspects of computer-based testing and try to come up with experimental designs that target these particular differences (i.e. different item type, the possibility of taking notes, etc.). The qualitative part of this dissertation should help to generate ideas for these more specific designs.

Overall, regarding the answers to the research questions:
1. The test scores slightly differ between test modes but only in one subtest (Destination) if a less conservative alpha level is observed. However, the difference is even smaller than in Study 1, here amounting to less than half a correct answer extra in the PPT mode.
2. Statistically significant gender differences with respect to the mode of test administration have not been observed, which confirms results from Study 1.
3. Some interesting outcomes are discovered when student performance on individual item types with respect to the two modes is statistically examined. While it is to be expected that short answer items will in general be more difficult for students than multiple-choice items (given the higher complexity of the short answer task type), which was confirmed, the statistically significant difference arrived at is that students score better on short answer items in the PPT mode than the CBT. This is only true for the Destination subtest.

Similarly to Study 1, the results of Study 2 did not bring enough evidence in favour of the effect of testing mode or gender on the students test scores, but did

report differences with respect to the item type and the mode of administration. This suggests that the effect of testing mode on people's performance is determined by the type of item (i.e. short answer item scores are lower when done in the CBT mode). This will be further explored using the qualitative data collected.

5.4 Pilot Qualitative Data Analysis

During the Pilot stage in 2014 (for details see 4.5), qualitative data was collected after each test mode intervention. First, the students were asked how they felt about the pencil and paper-based test mode, as that was the first test that they took. Second, after taking the computer-based test, they provided insights about the computer-based test mode as well. In the concrete, they were asked to write down one thing they liked and one thing they did not like about each mode of test administration. The data was then analysed thematically and coded, using an emergent coding system.

The following categories and themes emerged (the themes will always be capitalized for easier orientation):

5.4.1 Positives of the Pencil and Paper-based test

Pencil and paper-based test (PPT) advantages were mainly connected to Manipulation, Orientation and Attitudes. Regarding Manipulation and Orientation, according to students, the PPT was easier to fill in, navigate, and the danger of making typos was less likely. Students thus preferred writing by hand to typing. The theme Attitude was manifested through students feeling less stressed, more comfortable, natural and relaxed, or simply students having a better feeling. They enjoyed the friendly atmosphere and the human touch. Students also insisted that they pay greater attention with pencil and paper-based tests and are thus able to concentrate better since the technology does not distract them. Another very common theme was the appreciation of the PPT being marked by a human rater. Students considered that fairer and less intimidating. The themes of Familiarity with the PPT mode or Physicality (i.e. being able to touch the test) were also quite frequent.

5.4.2 Negatives of the Pencil and Paper-based test

Concerning the Pencil and Paper-based test (PPT) disadvantages, there were three common recurring themes, namely having to wait for the Results, which was the most common, Nothing (i.e. students could not think of anything that they did not like), and Handwriting. Students were mostly worried that their handwriting would not be legible, some thought about the teachers marking their tests and felt sorry for them having to decipher their answers. Some students felt that the PPT was bad for the Environment and others were unhappy about the Layout because they had problems fitting their responses into the small spaces provided. Changing Answers also proved problematic and students complained about the untidy and chaotic impression that it gives when they cross out some answers. In terms of Time, a few students felt the PPT mode took longer than the CBT and a few individuals missed the timer helping them check how much time is left.

5.4.3 Positives of the Computer-based test

Computer-based test advantages (CBT) mirrored the PPT disadvantages to a certain extent and the most frequent themes were Immediate Results, and other aspects related to technology, such as the ease of Changing Answers, seeing how much Time is left, or no problems with Handwriting or Legibility. Other themes included information about the Content of the test, happiness of students who passed, or the Nothing theme, in which some of the students claimed that they did not particularly like anything about the CBT. There were also comments grouped under the themes of Manipulation and Orientation and some students felt the CBT was quicker and found typing faster than writing by hand. As for some Attitudes or Feelings expressed, students appreciated that they could concentrate better, a few considered the CBT more comfortable and practical and some described themselves as Computer Fans and simply enjoyed the testing process more because it was computerized. The last stated point corresponds with the theoretical debate provided in 3.3.

5.4.4 Negatives of the Computer-based test

In terms of the Computer-based test (CBT) disadvantages, the Nothing or Don't know themes were the most common, which suggested that about a third of students did not have any major issues to share at that time. Other themes were often linked to Attitudes or Feelings, students felt more nervous, stressed, or found it harder to concentrate. For some, the CBT mode was less natural for them, some even claimed that they hate computers or consider the CBT mode less friendly and stricter. This goes together with scoring, which in the students' view is not to be trusted and they prefer the teachers to mark their tests. Some even mentioned seeing the Results immediately as a disadvantage. Other Technology-related themes included technical difficulties, complaints about the Screen, keyboards, problems with Manipulation and Orientation. They mentioned a fear of clicking on the wrong thing, and spelling incorrectly. They felt that this was due to either not being able to see mistakes as clearly on the screen as they could on paper or not being able to actually try out the spelling of an item on a piece of paper.

5.5 Study 1 Qualitative Data Analysis

In contrast to the Pilot, In Study 1 the qualitative component of the study was not collected anonymously in order to enable the researcher to match students' quantitative and qualitative data. All students were assured that all the data provided would remain confidential and no students voiced any concerns. As mentioned in the Participants description (see 4.3), there was a certain drop out of students because of the 6-week gap in between the intervention and thus the numbers differ slightly.

Students were asked two open-ended questions concerning their likes and dislikes of the two modes of test administration. (See Appendix 4B for the feedback form the students were given.) The table below shows the numbers of student answers, including multiple responses.

Question concerning	Number of students	Multiple responses	Total
PPT +	126	53	179
PPT -	126	19	145
CBT +	129	32	161
CBT -	129	17	146

Multiple responses refer to students who provided more than one answer.

5.5.1 Positives of the Pencil and Paper-based test

The three most common themes identified in Study 1 in terms of what students liked about the Pencil and Paper-based test (PPT) were Manipulation, No Technology and Orientation. Regarding Manipulation, which was often contrasted with the manipulation of the CBT version, space for making notes and the ease of turning pages and reviewing answers were the most appreciated. As Student 105/2015 puts it: *'it's much easier to turn between the pages than to click'* or Student 95/2015 claims: *'you can make notes, you can flick through easily'*. Various forms derived from the word 'easy' were present in most of the students' comments regarding Manipulation of the PPT. The themes of Manipulation and Orientation were often mentioned together. For example, Student 48/2015 insists that *'it's so much easier to find yourself in it, no scrolling, just simple page turning with PP*' or Student 133/2015 appreciated *'being able to turn pages and see the test as a whole'*. Student 18/2015 adds that *'it is easier to go back, paper is more transparent and visible'*. Another theme related to both Manipulation and Orientation is Physicality. For example, students liked the possibility of touching the test or claimed they liked the smell of paper (Students 123, 76/2015 respectively). Student 114/2015 summarizes it nicely: *'I felt more at ease with having a physical copy in my hand, I can make notes and returning to questions is easier'*. Such comments surprisingly frequently came from male students, which is in contrast to the debate on gender preferences regarding the use of technology presented in the theoretical part (see 3.5.1 for details).

As for No Technology, apart from general comments, such as *'I prefer PP because I don't have to worry about the computer'* (Student 25/2015), *'no computer problems'* (Student 41/2015), or *'I do not have to worry about internet connection or broken computers'* (Student 50/2015), various sub-themes could be traced, mainly to do with the computer screen (*'no staring at PC screen'* (51/2015), *'eyes not hurting because of the screen'* (127/2015), etc.), no complicated log in processes (31, 37/2015) and elimination of typos (17, 99/2015). The No Technology theme seems to be a criticism of the CBT mode rather than pinpointing the advantages of the PPT. Other themes included Nothing, Time, Attitude, Difficulty, Writing and Tradition. The theme Nothing is quite straightforward, some students left the space provided blank, others wrote '*nothing*', *'It [PP] was the same as C.B*' or '*no difference*' (e.g. Students 35, 65, 96, 137, 23/2015).

There was total agreement in the Time theme as all students referring to time in Study 1 felt that the PPT was shorter than the CBT, although they seem to be aware of the fact that both the tests are of exactly the same length. For example, Student 40/2015 noted: *'felt shorter, maybe just my imagination though'* or Student 52/2015 reported that *'it seemed shorter although it wasn't'*. Student 147/2015 directly related the time it took to the mode: *'it seemed shorter, probably because we were not staring at the screen - it was less exhausting'*. One student claimed to have been able to manage his time better (103/2015) and two students were happy about the absence of the timer, which they found stressful (7, 130/2015).
This goes hand in hand with the theme of Attitude. Once again, rather than commenting on the specifics of the PPT mode, they contrasted it with the CBT, claiming that the PPT was more natural for them (e.g. 75, 50/2015), less stressful (13/2015), less frightening (47/2015), or they were simply less nervous (139/2015). The theme Concentration could be linked to these feelings too, since when students were not nervous or stressed, they could concentrate better, which was how, for example, Students 3, 30 and 45/2015 felt about the PPT.

As far as Difficulty is concerned, all the students whose responses were coded under this theme were persuaded that the PPT version was easier than the CBT (e.g. 93, 101, 106, 111/2015). The last two themes, namely Writing and

Tradition will be discussed together because students are used to taking pencil and paper-based tests and are thus more accustomed to writing with a pen. As Student 122/2015 testified: *'I prefer this type of test because I'm used to it from high school'*. Some students stated that they found writing on paper more comfortable (e.g. 53/2015), others were not specific and just put down *'writing'*, *'writing with a pen'*, or *'I like writing on paper'* (61, 136, 89/2015 respectively) as something that they enjoy about the PPT. A few students also said that they liked everything about the PPT and that they simply preferred this mode (e.g. 57, 94, 134/2015).

5.5.2 Negatives of the Pencil and Paper-based test

The most recurring themes identified in students' responses to the question concerning their PPT dislikes were the following: No Results, Nothing, Changing Answers and Handwriting. The theme No Results was considered the biggest disadvantage of the PPT mode with adverbs referring to speed (or rather its lack), e.g. results are not provided 'immediately', 'right away', straightaway' or verbs, such as 'don't know', 'can't see' or 'have to wait', referring to No Results. Some students accompanied their statements by *'but it's not a problem'* or *'it's not such a big deal'* (Students 48, 70/2015 respectively), showing that they do not really mind not getting the results immediately. Within the Nothing theme, most students simply noted 'nothing', although a few added a positive remark, such as *'nothing, it was great'* (130/2015). The theme of Changing answers was mostly identified in connection with the difficulty of making clear corrections without the PPT looking messy and untidy. Student 15/2015 maintains that *'scratching wrong answers looks untidy'*, or Student 38/2015 dislikes the following: *'When I change my answer, everyone can see that I made a mistake'*. Quite a few students mention the issue of Changing Answers in relation to Handwriting, for example, Student 85/2015 says: *'it's harder to change my answers and make it legible'*, or Student 109/2015 complains: *'I may lose points because of writing'* and *'I couldn't rewrite mistakes easily'*. Other students panic about the human rater not being able to decipher their handwriting, or they do not like the fact that they need to focus on their handwriting

for it to remain legible (e.g. 105/2015). As Student 104/2015 puts it: *'I'm afraid my handwriting will have a bad effect on my results'.*

Other themes included Time, Content-related, and Environment. Student opinions were quite divided as far as the Time theme is concerned. Surprisingly, some students viewed the fact that the PPT version *'seemed shorter than the CB'* as a disadvantage (e.g. Student 138/2015). Student 18/2015 even stated: *'it seemed short, I was nervous a page was missing'.* On the other hand, a few felt that the PPT took longer (e.g. 67, 96/2015) and some missed the timer (13, 48/2015).
The Content-related theme could be linked to the length as students expressed dislikes concerning the lack of vocabulary tested (23/2015) or not enough pictures (118, 129/2015). Some also complained about ambiguous answers and one student thought that there was a mistake in the test (126/2015), which actually turned out to be the student's oversight. Regarding the theme Environment, students agreed that the PPT is 'bad for the environment', 'not-eco-friendly', 'destroying' or 'killing' trees (21, 47, 67, 96/2015).

The last themes to be mentioned here are Layout and Format. Complaints related to the Layout were mostly about spaces for answers being too small or small font used on the PPT (e.g. Students 4, 59, 72, 80, 87). In terms of the Format, students voiced their general preferences towards the CBT, saying either that they did not like *'the paper form'* (e.g. 137/2015) or directly stating that '*CB is better*' (82/2015). Some provided reasons for their general comments, such as *'I got used to CB here'* (71/2015), or *'CB would be better, it resembles how I study at home'* (95/2015).

5.5.3 Positives of the Computer-based test

The theme Results was by far the most frequently expressed advantage of the CBT by the students. Once again, the immediacy of the results was praised (e.g. 4, 7, 21, 29, 40, 46, 48, 51, 65, et al/2015). Three more themes were also very common, namely Content-related, Time and Changing Answers. Quite a lot of students commented on the content of the test and were happy with the vocabulary and

grammar questions it included, found it practical (e.g. 58/2015), claimed that it was fair as it contained what was covered in class (e.g. 143/2015) and appreciated that the pictures were in colour as opposed to the PPT mode (e.g. 87/2015). Some students appreciated the high number of production tasks – the short answer questions and stated that they preferred those to Multiple Choice Items (e.g. 35/2015).

The theme Time mostly contained comments relating to the timer and the length of the test, e.g. *'visible timer'*, *'it wasn't long'*, *'quicker than pp'* (Students 61, 22, 52/2015 respectively). Students also expressed some attitudes in connection with the timer or the perceived length, e.g. Student 18/2015 claimed that: *'There was a timer so I was not stressed about time'* and Student 13/2015 stated: *'It seemed short, I wasn't so nervous anymore'*. Some also mentioned the Time theme together with the theme of Changing Answers, e.g. Student 119/2015 says: *'It seems to be quicker, you don't have to scratch when you make a mistake*'. The possibility of changing answers multiple times, deleting instead of having to scratch over incorrect answers, and overall the ease of changing answers were what students mostly enjoyed (e.g. 35, 81, 14, 64, 103/2015). The theme of No Handwriting resurfaced relating to Changing Answers with students claiming that they *'can write legibly on a PC'* or that they *'do not need to worry about handwriting'* (e.g. Students 3, 109/2015 respectively).

Other themes included Nothing, Performance-related, Difficulty, Real Life Resemblance and General. The theme Nothing was pretty straightforward. This time, most of the students plainly stated *'nothing'* (e.g. 8, 12, 111/2015). The Performance-related and Difficulty themes can be presented together as students believed that the test was easier and their scores were higher compared to Achievement Test 1 (e.g. 10, 44, 49/2015). A few students liked the fact that there was *'no change of medium – study on computer, test on computer'* (e.g. 38/2015), and their comments were put under the theme Real Life Resemblance. In a more general sense, this theme could have been placed under the Format, too. Students 60 and 110/2015 were very positive concerning the CBT mode and stated that: *'I really like these tests on the computer'* and *'It's more entertaining, clicking and typing, and*

overall nicer'. These were labelled as computer fans. General comments were much less enthusiastic, though still put under the likes of the CBT mode, e.g. Student 74/2015 said: *'I'm getting used to them, so it's ok'*, or Student 86/2015 noted: *'It wasn't as bad as I thought'*.

5.5.4 Negatives of the Computer-based test

Students' dislikes regarding the CBT mode were more evenly distributed than in the sections discussed above. Two themes still dominated and these were Nothing and Content-related comments. This time, the space for the student's answer was quite often left empty and it is thus difficult to say whether there was really nothing that the students disliked about the CBT mode or whether they were not sure about their answer and therefore left the space empty. As for the Content-related theme, some students complained about confusing multiple-choice questions (e.g. 20, 101/2015), some were adamant that *'open questions should not be included in computer tests'* (e.g. 12/2015), and others felt there were not enough pictures (e.g. 16, 58/ 2015).

Concerning the themes directly related to technology, Technology in general, Screen, Typos and Log in were often commented on. *'Fear of breaking the computer'* (25/2015), *'risk of computer freezing'* (37/2015), or further unspecified *'technical problems'* (128/2015) were all grouped under the general Technology theme. The Screen theme was traced in the following complaints: *'bright computer screen, - sore eyes'*, *'my eyes get tired from staring at the screen'*, *'bad for my eyes'*, and *'the screen makes it harder to concentrate'* (6, 87, 115, 110/2015 respectively).

The next three themes to be discussed here are Manipulation, Orientation and Difficulty. The theme of Manipulation was mainly contrasted with manipulating the PPT mode and the students complained that it *'is impossible to underline, cross out in the CB test'* (e.g. Student 107/2015). Student 114/2015 agreed: *'I couldn't write notes /comments /whatever on the edge of the paper as I would on a paper test'* and Student 17/2015 missed writing *'helpful notes'*. Regarding Orientation, students felt that it was *'easy to overlook something'*, they could not *'see all questions together'*, or

that it was *'difficult to revise'* (91, 18, 41/2015 respectively). The theme of Difficulty was identified in comments, such as *'make the test easier'*, *'it was too hard'*, and *'grammar difficult'* (26, 35, 2/2015 respectively). In the researcher's view, these complaints stemmed from the fact that the students were shown their preliminary results straight after the CBT and knowing their score, they commented on the difficulty. That would explain why there were no such negative comments regarding the PPT mode.

Results, Performance-related, Time and Attitude are the last themes to be presented. In terms of Results, some students mentioned seeing the result as a disadvantage (e.g. 38, 68, 87/2015). This could be explained in the same way as with the complaints about the Difficulty mentioned above. The Performance-related theme falls under this category, too. Students might have been unhappy about their score and that is why they noted it down as something that they did not like about the CBT mode (e.g. 33, 122, 88, 68/2015). Interestingly, there were also voices questioning the scoring of the CBT mode, e.g. *'I don't know if the PC accepts the answer I give'* (102/2015), and quite a strong belief that *'only a person can evaluate a test properly, not a computer'* (121/2015). Student 114/2015 adds: *'the chance that I wrote a correct but unconventional answer is high, which the computer might not have anticipated'*. It is worth pointing out that these comments were provided by male students, who were thought to be greater technology fans than their female counterparts, at least according to the studies mentioned in the theoretical part (see 3.5.1). The main concerns about the theme of Time were the timer (stressful, disturbing, etc.) and not enough time (e.g. 36, 130/2015). Quite strong feelings were grouped under the theme Attitude. These included unhappiness, fear, hate, and stress as far as the CBT mode is concerned. Student 93/2015 makes a direct comparison of the CBT and PPT and insists: *'I was definitely more nervous compared to PP'*. Student 72/2015 states: 'I hate the computer, the ticking time is depressing'. These are quite strong emotions that were not expressed in connection with the PPT mode.

5.6 Study 2 Qualitative Data Analysis

Similarly to previous phases (Pilot and Study 1), all the data obtained from the students in Study 2 was coded using an emergent coding technique, and although most of the themes identified previously reappeared and were thus maintained, a few new ones had to be added. This suggested that conducting the qualitative component of Study 2 was definitely justified since the point of data saturation has not been reached and the information gained still led to further clarification of the research questions.

Students were asked the same open-ended questions concerning their likes and dislikes of the two modes of test administration, only this time all at once as they took both the modes on one day. (See Appendix 4C for the feedback form.)

The chart below details the numbers of student answers, including multiple responses.

Question concerning	Number of students	Multiple responses	Total
PPT +	126	28	154
PPT -	126	7	133
CBT +	126	14	140
CBT -	126	22	148

Multiple responses refer to students who provided more than one answer.

The chart does not aim to quantify the qualitative data received but rather to point to the differences in how much students comment on the individual aspects, which could indicate how strongly they feel or care about the mode. We can see that compared to Study 1, students in Study 2 provide considerably fewer multiple responses, which might suggest that they do not feel so strongly about the different modes of administration anymore.

5.6.1 Positives of the Pencil and Paper-based test

Among the most often voiced likes regarding the pencil and paper-based tests (PPT), the following major themes were identified: Manipulation, Time, Tradition, Attitude and No Technology. The most recurrent theme of Manipulation contained information on making notes, the possibility of crossing out wrong answers, skipping exercises, and checking previously answered questions easily. Students particularly enjoyed the opportunity to make side notes, some of them even stated that making notes helped them think more, which links it to the Performance-related theme. For example, Student 15/2016 claims: *'I was able to write side notes and think about the options more and it was easy to return to questions I was unsure about.'* Or as student 71/2016 puts it: '*I liked that I could write on it – try out different spelling, write both possibilities, compare and decide later*.' Manipulation was sometimes mentioned in connection to yet another theme, i.e. Orientation. Student 37/2016 summarized it aptly when saying: *'I liked that I could mark the points where I wasn't sure. It was easier to navigate. And I can cross out answers which are incorrect.'*

Within the theme of Time, students most often commented on the perceived length of the PPT, starting on time and not being rushed. They considered the pencil and paper-based test shorter, quicker, claimed that there were no delays and felt that they had more time to finish. This could also be linked to the theme identified as Difficulty, for which the students unanimously maintained that the PPT seemed easier than the CBT. This was confirmed by the analysis of the quantitative data, although the differences were not significant.

As for Tradition, students mentioned that they are used to this mode because all of their high school tests were in paper form. They also enjoyed the spirit of tradition and described the PPT as regular or classic. Student 123/2016 added: *'I don't like changes and since my whole life I've been taking paper tests, it was nice to do it again.'* Some students also mentioned the theme of comfort in relation to Tradition as, for example, student 110/2016's quote testifies: *'I'm used to writing tests on paper so I would say it is much more comfortable for me.'*

The next commonly expressed area included various feelings students associated with the PPT, which were classified under the theme of Attitude. Here, the students stated that compared to the CBT, when taking the PPT they were less nervous or stressed, they felt more confident or relaxed and considered the PPT to be more private and personal. Student 123/2016 offers the following explanation: *'I don't know what but there's something about PP atmosphere that makes me like it. It's more personal and transparent.'* The word 'natural', which could perhaps be linked to the theme of Tradition discussed above, was also quite frequent in their descriptions. Interestingly, student 89/2016 believes that the PPT is more serious than the CBT and approaches it with more respect.

In terms of the No Technology theme, students mostly appreciated the absence of the screen, no technical difficulties, or no distracting noises (e.g. typing on the keyboard, or mouse clicking). For example, student 87/2016 confirmed this when saying: *'It didn't hurt my eyes. It is refreshing not to have to look at the screen for once.'* Some students also mentioned the advantage of not having to remember any passwords, which they often mistyped when stressed in the test situation. A few students also stated that they liked the fact that they did not get to see the results immediately, which could be placed under the No Technology theme, although it was given a separate code NR, i.e. No Results.

One category that should not be omitted here, as it was also quite a frequent one, was 'Nothing'. Some students simply did not provide an answer regarding what they liked about the PPT or said 'nothing', 'I can't think of anything', or as Student 93/2016 put it: *'I'd really like to give you a helpful answers but there's no such thing about both the forms.'* This response could mean that Student 93/2016 either does not care which mode of test she takes or does not like the achievement tests in general.

There were a few themes that were quite exceptional but worth mentioning. These include Performance-related, Physicality, Writing and General themes. The Performance-related theme was briefly mentioned above in relation to Manipulation. However, there were also quite a few students who believed that the PPT helped them to perform better, for example Student 1/2016 said: *'It gives me an*

impression that I'll get the answer right'. Or as Student 88/2016 testified: *'The written test somehow helps me to think better'*. Other accounts stated: *'By writing the word, I can remember and recall it easier'* (Student 98/2016) or *'I feel like what I'm thinking flows easier when I write myself'* (Student 113/2016). This would suggest that cognitive processing does differ depending on the mode of the test. Physicality, i.e. touching the test, can be an important factor too, as Student 63/2016 indicates: *'I like to be able to hold it in my hands'*. The theme of Writing goes hand in hand with the Physicality theme and students are happy about writing by hand, using a pen, and they prefer it to typing, which will be later demonstrated. The last General theme to be presented here contains some non-specific comments on good atmosphere, pleasant supervisors, or just a statement of general preference (i.e. Student 74/2016 saying *'I prefer PP to computer'* or Student 36/2016 claiming *'Everything, I prefer PP.'*)

5.6.2 Negatives of the Pencil and Paper-based test

The dislikes the students have commented on in terms of the Pencil and Paper-based test (PPT) mainly contained the themes of Nothing, Classroom, No Results, Time, Changing Answers, and Handwriting. The theme Nothing, meaning that there is nothing that they did not like about the PPT mode, has largely been the most frequent one. It represented almost a third of all the answers provided. Unfortunately, another theme, which was very commonly voiced, was caused by external factors not directly related to the mode of the test, namely lack of space and uncomfortable classroom furniture. This theme was entitled Classroom and the students stated the following: *'The space provided wasn't big enough, the tables were really small.'* (Student 10/2016), or *'The chairs with desks attached were unpleasant'* (Student 16/2016). The word 'uncomfortable' in connection to chairs and desks was often repeated (e.g. Students 12, 52, 72, 88, 108 and 114/ 2016). In relation to space, Cheating was identified as another theme. For example, Student 115/2016 stated: *'It's easy to cheat for some of these people here'*, where 'here' refers to the classroom. The No Results theme was quite straightforward and self-explanatory.

Students complained about not knowing the results immediately or having to wait for them.

In terms of Time, students had different concerns compared to the positives identified above. They missed the timer, which is an inherent part of the CBT and felt '*It was not clear when the test was ending*' or how much time (Student 30/2016). Others commented on having to wait a long time between the two test administrations (e.g. Students 40 or 51/2016), which in the researcher's view does not relate to the PPT directly. A few students mentioned the exact reverse of what was claimed regarding Time, stating that the PPT was longer. For example, Student 42/2016 felt the following: *'Personally, it takes me longer to finish a pp test.'* The length was sometimes explicitly mentioned in connection to the theme of Handwriting, as Student 58/2016 claims: *'It's more time consuming for me to write by hand, I have to focus on handwriting otherwise it's illegible for the teacher.'* Legibility was indeed the main concern within the Handwriting theme. Some students even thought about the teachers having difficulties when correcting their tests: '*Students scribbling is hard for the teachers to mark*' (Student 110/2016) or '*my handwriting is sometimes illegible, which might be a problem for the person correcting my test*' (Student 44/2016). Student 113/2016 even expressed a fear of not having an answer recognized because of bad handwriting.

As far as Changing Answers is concerned, students did not like crossing out and changing their answers as they considered it messy and more complicated than in the CBT mode. Student 46/2016 simply said: *'I can't change answers easily'* or Student 61/2016 added: *'I don't like that I'm able to write correct answers only once'*. Student 1/2016 claimed that *'crossing out answers is messy but it's only aesthetical problem'*. This suggests that it was not a major issue for Student 1/2016, while for Student 71/2016, who complained that when changing answers, *'everyone can see what mistakes I have done',* it seemed to be more of a problem. Some emotionally charged words, such as *'it looks terrible when you want to correct yourself'* (Student 119/2016) were also used.

There were other themes, such as Content-related, Difficulty, Orientation, and General but these were rather rare. As for the Content-related theme, pictures

were criticized for being black and white or simply worse quality than in the CBT mode (e.g. Students 66, 76/2016) and one student (54/2016) felt that the test contained different tasks from the ones that they studied for. Some students mentioned that this test was more difficult than the previous achievement test (e.g. 79/2016). Concerning the Orientation theme, Student 50/2016 missed *'a lovely little table telling you which answers are already answered'*, which is inherent in the CBT mode. The General theme contained a few complaints about having to write two tests instead of one (Students 19, 25, and 78/2016) and one note aimed at the invigilators: *'teachers were talking, opening the door, which was disturbing'* (Student 69/2016).

5.6.3 Positives of the Computer-based test

There were three main areas that students expressed their likes in regarding the Computer-based test (CBT), namely Results, Time, and Nothing. Once again, the students enjoyed the instant availability of results, which, as discussed in the theoretical part, is a great advantage of a CBT. Students are always told that their score is not final and that a human rater will go through the answers of the open-ended question types and some showed their awareness of this in their contributions: *'you get an approximate score immediately'* (Student 113/2016). The Time theme can be divided into two sub-themes, one concerns the presence of the timer in the CBT mode and the other concerns the length of the test. Most of the students commented on the benefits of the timer. Student 56/2016 says: *'I liked that we could see the timer, so we know exactly how long we will have to suffer'* and Student 15/2016 simply states: *'I'm able to check how much time is left at any time'*. A few students felt that the CBT was faster than PPT but some were not sure, for example, Student 97/2016 just put *'Faster than writing?',* which does not sound very persuasive as an advantage of the CBT mode.

The theme Nothing was also common, some students left the space for the answer completely blank, others claimed 'Nothing', 'I didn't like it' or 'I'm not a fan of computers, so nothing' (Student 73, 60, 62/2016), etc. Other themes included

Changing Answers, Comfort and Attitude. In terms of Changing answers, this theme can be seen as reversed from the PPT dislikes and turned into a CBT likes. It is interesting to point out that different students voiced it as positives of CBT. For example, Student 53/2016 believed that *'it's easier to correct my answers when I change my mind'*, or as Student 107/2016 put it: *'one click and my answer is changed'*. Student 55/2016 remained somewhat hesitant whether changing answers was always beneficial when claiming: *'I could correct the answer several times (it's not always a good thing though)'*.

The theme of comfort was represented by students' straightforward claims that they find the test more comfortable. Almost none of them specified why they find the CBT mode more comfortable except for Student 88/2016 who said: *'it is more comfortable for people like me who can't write very well as you can see'*. The feedback was handwritten, so that is what they were referring to in the second part of the quote. Attitude was a slightly tricky theme as students expressed their laziness, for example, and thus the suitability of the CBT mode (Student 81/2016), found the test to be more practical (Students 85, 94/2016) or felt more relaxed (Student 89/2016).

The following themes were mentioned by individuals but are rather important for grasping the overall impressions. The theme Real Life Resemblance was demonstrated by Student 58/2016 when saying: *'It's much better, I write everything on computer'* and Student 80/2016 claiming: *'I learn from my computer, so it's easier for me to remember everything'*. Orientation, which was also identified as something that the students liked in the PPT mode, was referred to in a few cases, such as *'I can see which questions I've already answered easily'*, or *'It's more přehledný (sorry I don't know the English word)'* (Students 37, 39/2016). One student commented on the theme of Environment, being pleased that 'we don't waste paper' (Student 4/2016) and a few students were identified as Computer fans, in accordance with Brown's beliefs that students might find the testing process more enjoyable when computerized as discussed in the theoretical part (see 3.3), because they said that they liked working with computers or found it 'fun to fill out the answers' (Students 105, 123/2016). In terms of Difficulty, a few students

considered the CBT version easier (Students 13, 52, 79 and 121/2016). Students 9 and 65/2016 appreciated the classroom layout and comfortable sitting respectively, which was probably contrasted with the classroom in which the PPT version took place. Student 86/2016 maintained that she found the CBT 'easier to concentrate'.

5.6.4 Negatives of the Computer-based test

The most commonly expressed negatives of the CBT were categorized under the following themes: Technology-related, Nothing, and Time. The most recurrent theme was Technology-related, which can be further divided into Technology (in general), Typos, Noise, Log in, and Screen. As far as Technology in general is concerned, there were voices, such as *'it can be problematic'* (Student 4/2016), *'the computer was very slow'* (Student 13/2016) or comments that partially express the students' attitudes towards the CBT mode too, e.g. *'I don't trust machines'* (Student 64/2016) or *'I'm scared of technical difficulties'* (Student 63/2016). Some Erasmus students expressed issues with the setting of the 'qwertz' setting of the keyboard or the keyboard in general (e.g. Student 113/2016).

The theme Typos was also very common and students believed that they are more prone to making typos in the CBT mode. For example, Student 89/2016 claims: *'I'm bad with typos, when I write on paper I have more control'* and Student 108/2016 says: *'There is a high risk of typos, I have to double-check everything'*. Student 43/2016 adds that *'you make more mistakes because of writing on keyboard'*. I would like to mention one more theme here identified as falling under Mistakes but fitting within the Technology realm, too. Students were alarmed that only one 'misclick' may result in the answer being wrong, as Student 26/2016 puts it: '*it is more possible to make a mistake by accidentally clicking on the wrong answer*'. Another theme directly related to technology is Noise produced by the computers. Mostly students are disturbed by 'clicking noises' or 'typing on the keyboard' (e.g. Students 8, 21, 59, 69, 75/2016). These can be linked to some negative attitudes, as Student 120/2016's comment demonstrates: *'When other people are typing, it's annoying and distractful'*.

In terms of Log in, students mainly complain about *'a complicated system of logging in and out'* (Student 42/2016), *'the time spent logging in and out'* (Student 105/2016) and once again, the whole process can be linked to negative attitudes, e.g. *'logging into the computers is complicated and confusing, it stressed us'* (Student 118/2016). In this day and age the theme Screen seems rather surprising, given how much time students spend in front of the computer. However, students still voiced their unhappiness stating *'it's not a pleasure for my eyes'*, *'the screen is too bright'*, *'my eyes and head hurt'*, which could possibly have an impact on concentration, or simply *'unhealthy for eyes'* (Students 57, 81, 80, 87/2016 respectively et al.).

The second most often repeated theme identified was Nothing. Similarly to the Nothing theme discussed above, students either left the space for a negative comment about the CBT mode blank or stated *'nothing'*, *'nothing in particular'* or *'nothing, I liked everything about it'* (e.g. Students 71, 32, 102/2016). Concerning the theme Time, there is a bit of a discrepancy between the students' answers, some claiming what they did not like was a delayed start (e.g. Students 11, 19, 23/ 2016) and others complaining that they had to wait for a long time at the end of the test and were not allowed to leave the room until everybody finished (e.g. 16, 30, 121/ 2016). This can be attributed to the different groups these somewhat opposing views originate from. One of the groups was indeed delayed because of some technical difficulties with logging in, while the other group started on time, which probably left some of the students thinking that they had more time at the end of the test. Another interesting difference in opinions seems to be brought about by the timer – while some complain about the timer disturbing them, find it stressful, which can again be linked to attitudes (e.g. Students 39, 65, 84, 85/2016), others complain about not being able to see the timer when they scroll down (e.g. Student 116/2016).

Other themes included Results, and previously touched upon Attitude, Difficulty, and Manipulation. Interestingly, in contrast to the students who saw getting the results straight after the CBT as an undisputable advantage, there are quite a few students who would have preferred not to know their results

immediately. Two students even mentioned the Result theme under both positives and negatives of the CBT (111, 114/2016), seeing it as positive as well as negative. The researcher is inclined to think that primarily the students who were unhappy about their result mentioned it as something that they did not like about the CBT mode (e.g. 5, 44, 79, 92, 96/2016). One student complained about the lack of feedback provided saying *'I was surprised to see my grade and didn't know where I missed out'* (60/2016), which is an area that should definitely be paid attention to as the theoretical part demonstrates (see 2.2.6 and 2.3.3 for details). A few students also commented on the scoring, one saying *'I don't believe the computer scoring, I dread it rules out a question that is correct just in different form, somebody always has to check manually, so it does not speed the process of grading so much'* (99/2016), and another thought that teachers would be more lenient in their corrections: *'I would get extra points when marking is done by teachers I think, computer will only take answers as it 'knows' them'* (110/2016).

The theme Attitude was identified in the following examples: *'I don't like tests on computer, they make me nervous'* (45/2016), *'nothing specific, it's just a feeling'* (67/2016), and *'I felt more nervous'* (117/2016). The most common adjectives used within this theme were 'nervous' and 'stressed'. There were students who linked their negative feelings regarding the CBT mode to the inability to concentrate, e.g. Student 88/2016 stated: *'I don't like it as I feel distracted while taking the PC test'*. Student 9/2016 added: *'I can't fully focus on computer based tests'*.

In terms of Difficulty, students expressed their views unanimously this time, stating that they found the CBT harder than its PPT counterpart (e.g. 37, 58, 90/2016) and in terms of Manipulation, students had issues with scrolling, having to skip pages and no opportunity to make notes (e.g. 2, 36, 66, 101/2016).

5.7 Research Tool Development

As mentioned in Chapter 4 detailing Research Methodology, the researcher has also devised and piloted a quantitative questionnaire, which is based on the qualitative data gained directly from the students. The reason for that is two-fold.

First, in accordance with Dörnyei (2007), who considers questionnaires to be a versatile and time-efficient tool gathering a lot of information quickly, it is thought that such an online questionnaire can be easily administered to students in our context in the future and data regarding the two modes of test administration can be obtained more efficiently. Second, while there are various validated questionnaires regarding computer familiarity, usage and anxiety (see below), none of those is directly linked to language testing and comparing the two modes of test administration. It is thus believed that the new tool can serve as a research instrument for institutions, which either already use CBTs and want to find out how their students feel about them, or are considering implementing computer-based testing.

The questionnaire is based on three previously validated instruments, namely an instrument used in the Program for International Student Assessment (PISA) investigating computer familiarity, Knezek and Christensen's Computer Attitude Questionnaire (1997) and the last instrument being the Computer Familiarity Questionnaire devised by Weir et al. (2004). It was first compiled for the author's master thesis research dealing with the same topic, which was carried out in 2006. The present questionnaire has been updated and revised in accordance with Noyes and Garland's (2008) claim against using, for example, computer attitude scales devised in the 1990s because of possible changes in the construct as discussed in the theoretical part (3.5.1). The aim was to make sure that it corresponds to the latest developments in the area under scrutiny. Questions on Computer Usage have thus been mostly eliminated since the students work with computers on a daily basis, yet a number of questions related to the specifics of taking tests on computers have been added. Most of these came from the qualitative data collected during the Pilot in 2014 and some items were also added or reformulated after Study 1. Once the qualitative analysis was carried out, the main themes identified were put into categories and turned into statements.

When devising the questionnaire, Brown's (2001) guidelines for writing good survey questions were strictly adhered to, taking into consideration the form (length, clarity, no negative questions, no overlapping choices, etc.), the meaning

(the avoidance of double-barrelled questions, loaded words, leading, biased, embarrassing questions, etc.) and the respondents (language level, avoiding irrelevant questions, superfluous information, etc.). (For more details see Brown, pp. 45-55.)

Most of the questions make use of a closed response format with either alternative answers or Likert scale. In such multi-item scales Dörnyei (2007) stresses the importance of wording when asking about respondents' attitudes, feelings, etc. (p.103) and similarly to Brown, provides a list of rules about item wording, which was also observed when redesigning the questionnaire (p.108). Furthermore, there are a few open-ended questions asking the students to fill in some information based on their experience, which fall under the category of open-ended, i.e. qualitative data (Creswell 2015).

Table 20 presents the items that are designed to elicit data for the following indices. The indices and the specifications are based on the qualitative data gained and the areas covered in the theoretical part of the dissertation:

Index	Specification	Questions
Perceived ability	General	Q1
	Manipulation/ Orientation	Q8, Q9, Q11, Q12
	Typing/ Handwriting	Q20, Q22
Preference	Comfort/ Affect	Q2, Q4, Q6, Q10
	Manipulation/ Orientation	Q7, Q16, Q17, Q18
	Typing/ Handwriting	Q19
	Scoring	Q21, Q23
	Interest	Q13, Q24

Table 20: Questionnaire Indices

For more details concerning the questionnaire, the exact wording of the questions and a preliminary analysis, which includes the tool's overall reliability

estimates, the frequency of data obtained, and mean item scores and correlations with the whole reliability index divided into the two scales, i.e. Perceived ability and Preference, see Appendix 0 for a link to the storage drive.

Overall, the questionnaire results agree with the qualitative data collected, although the analysis is far from complete and will be further worked with. The final phase of the questionnaire development has not been reached yet, however, it is hoped that once fully validated, the tool will serve the purpose of efficiently collecting multiple data regarding student attitudes to the two modes of test administration.

6 MERGING THE DATA: FINDINGS AND DISCUSSION

In this chapter, the quantitative and qualitative results will be brought together, summarized, integrated and interpreted in order to answer the research questions.

6.1 Research Question 1

RQ 1. Are there significant differences in the scores from the Computer-based test (CBT) and Pencil and Paper-based test (PPT) modes?

As the analysis of the quantitative test data showed (see 5.1-5.3), it can be concluded that there are **no significant differences** in the scores gained from the CBT and PPT modes and the tests can be used interchangeably. Students in all three stages of the research, namely Pilot, Study 1 and Study 2, overall performed slightly better in the PPT mode than in the CBT but the differences are not statistically significant if the p value is kept at a conservative level alpha < .01. One subtest in Study 1 (Vocabulary) and one subtest in Study 2 (Destination) were marginally statistically different if the p value is kept at a more liberal level of alpha < .05. However, the difference would be less than half an extra correct answer per subtest and is thus negligible. These results are in agreement with Mead and Drasgow's meta-analysis of 28 studies concerning the mode of delivery discussed in the theoretical part, which reports that students mostly perform better in the PPT mode, however the differences are not significant (see 3.5 for details).

6.2 Research Question 2

RQ 2. Do the scores from the CBT and PPT modes differ in terms of gender?

Based on the discussion presented in the theoretical part (see 3.5.1 for details), it was expected that male students would perform better in the CBT mode and female students would gain better results in the PPT mode. This was not the case as both genders performed better in the PPT mode. Statistically, **no significant differences** were found either in Study 1 or Study 2 in connection with gender.

Nevertheless, this will be further examined when discussing student preferences in RQ 5 and RQ 6.

6.3 Research Question 3

RQ 3. Do the scores from the CBT and PPT modes differ with respect to the question type?

Special attention was paid to this in Study 2, because of some mixed results encountered in Study 1, which were not possible to investigate further as the versions were not identical. In Study 2, the researcher made sure that this aspect could be observed and **some significant differences** were indeed discovered. The conclusions arrived at are as follows:

- Students perform better in multiple choice questions as opposed to short answer items in both modes of test administration (this is to be expected given that multiple choice questions only test recognition knowledge while short answer questions test production and are thus more challenging for the students – for more details regarding the individual item types see 3.2.9).
- In terms of short answer questions, students perform better in the PPT mode than in the CBT mode. There was a highly significant difference in the Destination subtest of Study 2. This confirms the results of some of the studies discussed in the theoretical part (see 3.5), however the reason why the difference was significant only in one of the two subtests remains unclear.

6.4 Research Question 4

RQ 4. What are the advantages and disadvantages of the two modes of administration as viewed by students?

If the qualitative feedback data presented in the previous chapter is approached quantitatively, certain patterns start to emerge and some shifts in students' perceptions of the two modes over time become apparent. These will be

discussed now in order to answer RQ 4 and possibly serve as an explanation to some of the quantitative data discussed above and in 5.1-5.3.

The advantages of the PPT mode and the disadvantages of the CBT mode will be conferred together as they very often correlate as do the advantages of CBT and the disadvantages of PPT in the subsequent section.

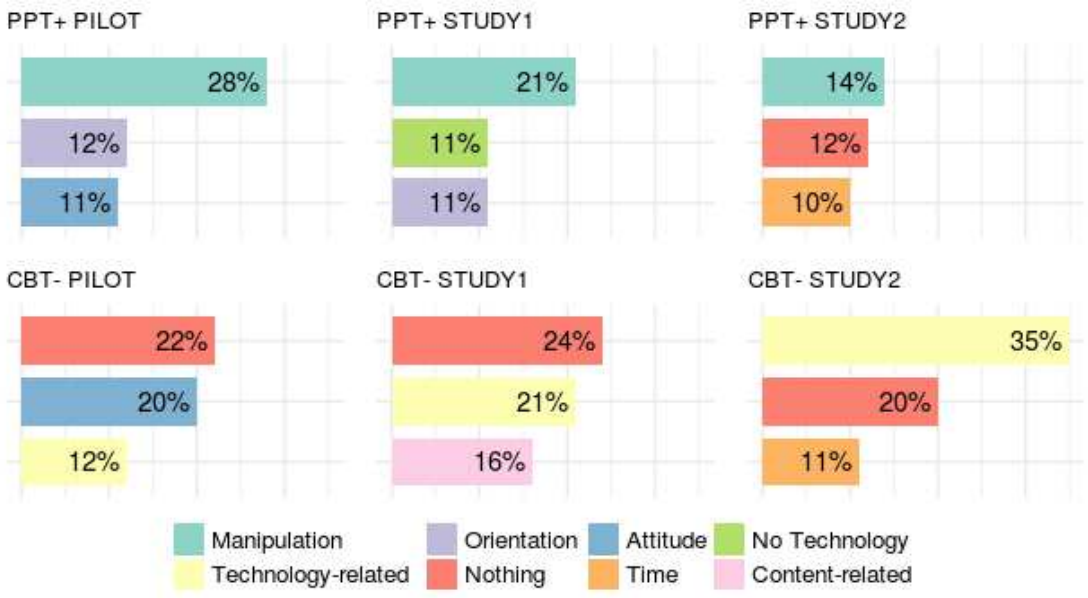

Figure 13: Advantages of the PPT mode (PPT+) and Disadvantages of the CBT mode (CBT-)

As Figure 13 demonstrates, the main advantages of the PPT mode as identified by the students slightly change over time, however, the theme Manipulation remains the most recurring one over the course of the research. It is mainly the possibility of making notes, trying out spelling, and crossing out incorrect options, which the students appreciate the most. Quite a few also believe that writing by hand helps them think. This could definitely be linked to the different performance on short answer versus multiple choice questions discussed in RQ3.

While in the Pilot and Study 1, the theme of Orientation was among the most frequently cited advantages of PPTs, in Study 2, the majority of students no longer

felt that way. This could be attributed to the fact that students have become more used to working with computers in their academic lives and not seeing the test as a whole or difficulties with navigating through the CBT mode are not seen as a threat anymore. Nevertheless, students still complain about technology-related issues, such as technical difficulties, log in processes, having to stare at the screen and a higher risk of typos, which is apparent even in their comments about the PPT mode when they state that one of its main advantages is no technology.

It seems that students have become more content with and relaxed about the CBT mode over time, as the frequently cited Attitude theme concerning positive feelings towards the PPT mode and negative feelings towards the CBT mode, voiced especially in the Pilot, has ceased to be of primary concern. The Time theme is reflected in both modes. In the PPT mode, students feel the test takes a short time and is quicker, while in the CBT mode, they maintain that it takes a longer time because of technical delays, etc. The two nicely complement and confirm one another in Study 2. As for the Content-related theme present in Study 1, complaints were mainly directed at the amount of material that the students were required to study for the test. It suggests that these students did not find any disadvantages related to the CBT mode and thus commented on the content. Finally, it needs to be stated that the theme Nothing was a very frequent one in all the stages of the research. This definitely points to the fact that there is a high number of students who do not have any problems with the CBT mode.

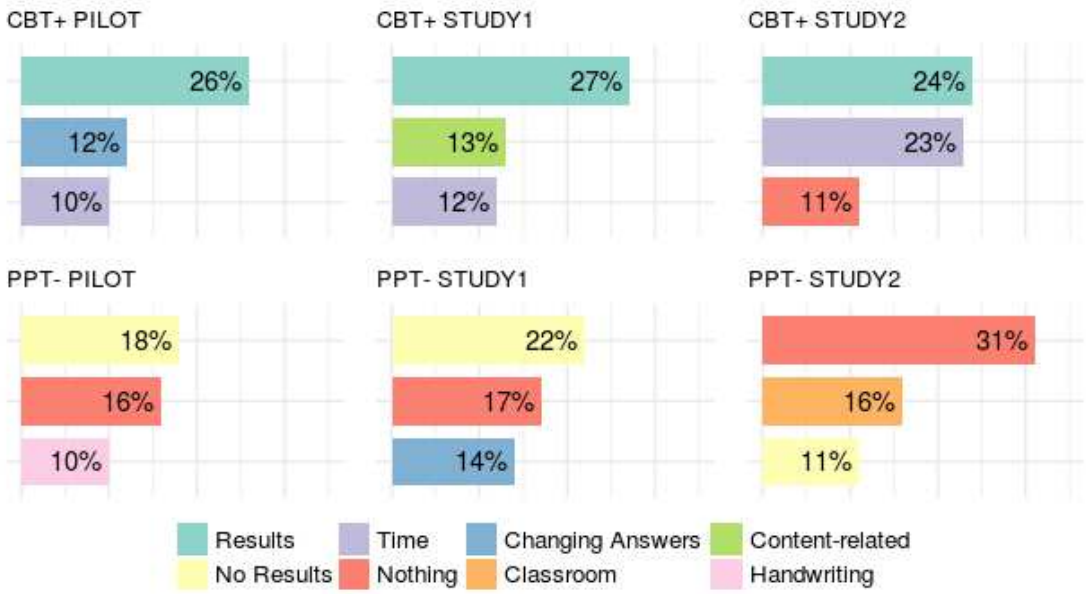

Figure 14: Advantages of the CBT mode (CBT+) and Disadvantages of the PPT mode (PPT-)

The main advantage of the CBT mode as identified by the students over the three years is indisputably the immediacy of results, which is in agreement with, for example, Noyes and Garland's (2008) claims presented in 3.3. This is further supported by the most commonly expressed disadvantage of the PPT mode, i.e. the lack of immediate results. It is interesting to point out that Study 2 was the only one in which the No Results theme was not viewed as the biggest disadvantage of the PPT mode. Changing answers was considered easy and tidy and represented the second most commonly voiced positive of the CBT in the Pilot. Subsequently, in Study 1, Changing answers frequently appeared as a negative of the PPT mode. The theme Time was present in all three stages and students mainly appreciated the possibility to check how much time is left in the CBT mode. As for the Content-related theme, students were positive about the test content of the CBT mode in Study 1, although about the same number of students in Study 1 were critical of the same test and mode (as mentioned above). Interestingly, in the PPT mode, students seldom commented on the content of the tests. There is no clear explanation for this except for the possible reason mentioned above, i.e. that students felt that they had nothing to say about the mode and thus focused on the test content. Concerning the PPT disadvantages, once again the theme Nothing reappeared in all the stages of the

research, which suggests that quite a lot of students do not have any major issues with the PPT mode either. The theme Classroom is slightly unfortunate as the only room available for the PPT mode in Study 2 was far from ideal. Students only had small desks attached to uncomfortable chairs at their disposal and the space was not big enough and that is why they expressed their unhappiness with the classroom setting. This was not the case in the Pilot or Study 1, hence there were no comments related to the Classroom. It is a shame but at the same time it reveals how much classroom environment influences the students.

6.5 Research Question 5

RQ 5. What are the students' preferences concerning the two modes of administration?

In Study 1 and Study 2, all students were asked to state their preferences regarding the two modes in the feedback forms administered immediately after the test mode interventions.

In **Study 1** with the total of 114 research participants (71 females and 43 males), 41.2% (n=47) were in favour of the **pencil and paper-based test mode**, 39.5% (n=45) preferred computer-based tests and 19.3% (n=22) stated no preference. The preferences were thus quite equally divided between the two modes. It is interesting to see the breakdown of the students' answers by gender.

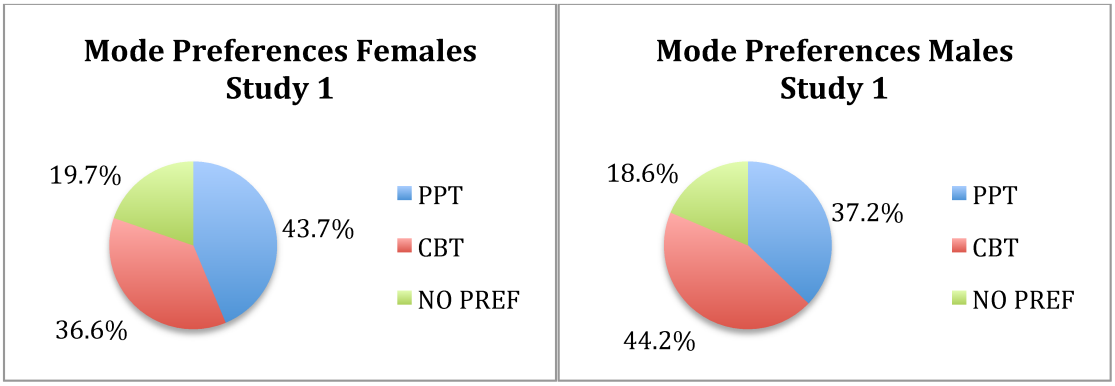

Figure 15: Study 1 Test Mode preferences by gender

As Figure 15 demonstrates, females (n=71) prefer PPTs, while males prefer CBTs in Study 1. Surprisingly, although the female/ male proportion differs, the percentages are as if reversed, only the 'no preference' groups remain unchanged.

In **Study 2** with the total of 126 research participants (86 females and 40 males), the shift in the student preferences or rather in the popularity of the 'no preference' category becomes apparent. The **'no preference'** category is chosen by 37.3% (n=47), followed by computer-based test preference with 34.1% (n=43) and pencil and paper-based test preference with 28.6% (n=36). Preferences according to gender are as follows:

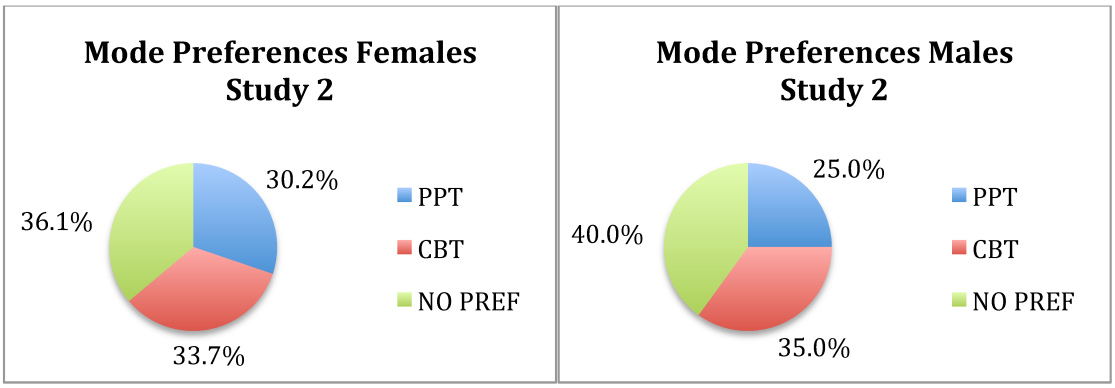

Figure 16: Study 2 Test Mode preferences by gender

Figure 16 shows that the CBT mode is the preferred choice over the PPT mode by both females and males, although the difference is more pronounced with males. Nevertheless, the 'no preference' category remains the most popular, which signals that students' opinions towards the two modes have become less extreme.

When looking at the students' test results, in Study 1 out of 114 participants 54.4% (n=62) performed better in the PPT mode, 41.2% (n=47) had better scores in the CBT mode and 4.4% (n=5) achieved exactly the same score in both modes. Interestingly, in contrast to the preferences stated above (see Figure 16), males

performed noticeably better in the PPT mode, although they preferred the CBT mode. See the breakdown of results below.

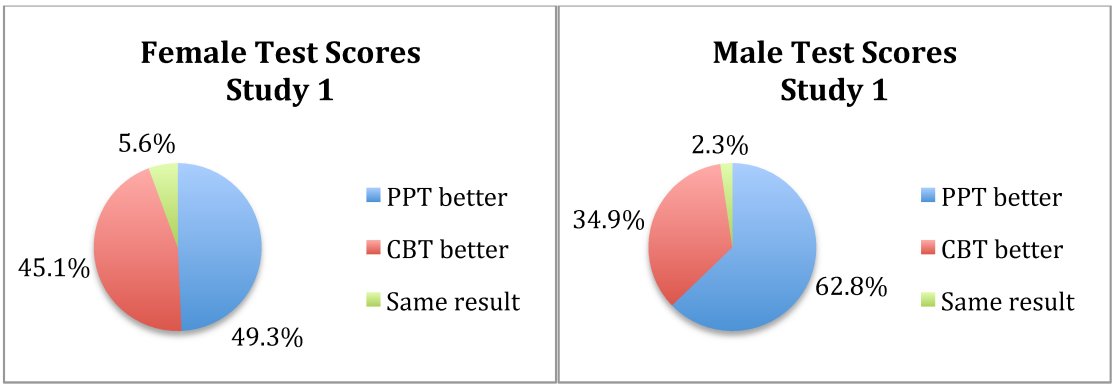

Figure 17: Study 1 Test scores by gender

Similarly, in Study 2, out of 126 participants 56.3% (n=71) achieved better scores in the PPT mode, 40.5% (n=51) performed better in the CBT mode and 3.2% (n=4) had identical scores in both modes. This time, the differences between the student preferences and their scores were even more pronounced with respect to both genders.

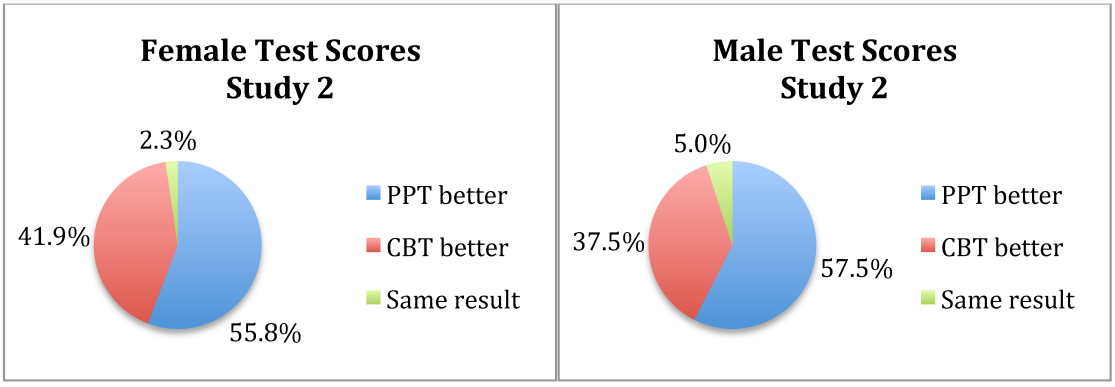

Figure 18: Study 2 Test scores by gender

Figure 18 demonstrates that in Study 2 a higher percentage of both females and males perform better in the PPT mode, although their preferences are in contrast to this. However, it should be noted that the exact test scores (i.e. how big

the difference in the scores is) are not specified here. (For details regarding the test scores, see 5.3.2 in the Quantitative Test Data Analysis section).

6.6 Research Question 6

RQ 6. Is there any clear link between the student preferences and the results they gain?

In order to answer RQ 6, qualitative and quantitative test data from the same individuals will now be merged to see whether students' preferences regarding the two modes of administration (as expressed in their feedback forms) affect their test scores gained statistically.

The statistical model employed makes use of both quantitative and qualitative data and seeks to ascertain whether there are statistically significant differences between the mean subtest scores in connection to student preferences. Similarly to statistical models in 5.2.3 and 5.3.3, a mixed-effect model was constructed, this time with fixed effects of *testing mode (CBT vs PPT) and preference (CBT, PPT, no preference) and random effects of version (1 and 2) and test-taker (student id).*

As far as **Study 1** is concerned, **no significant results** were found in either of the subtests. Likewise, the interaction between testing mode and preference was not significant, which suggests no effect of preference on test scores. For detailed results, see Appendix 5A. Regarding **Study 2**, the Vocabulary subtest did not show any significant results either, while the Destination subtest was **marginally significant** in terms of interaction between testing mode and preference (i.e. beta = 1.2, SE = 0.55, p = 0.032 and beta = 1.23, SE = 0.59, p = 0.038 for "no preference" and "preference for PPT", respectively). Similarly to other statistical models discussed in 5.3.3, students who prefer the PPT mode or have no preference would gain slightly better results when writing tests in the PPT mode. For detailed results, see Appendix 5B. The magnitude of the effect is 1.2 points, which amounts to almost 1.5 items extra correct in the PPT mode as opposed to the CBT mode. This is in agreement with the better PPT test scores gained by the students irrespective of their preferences. However, once again, the result is statistically significant only if

the p value is kept at a more liberal level of alpha < .05 and in only one out of four subtests.

It can thus be concluded that overall **no clear link between student preferences and their test scores** has been documented and student preferences therefore do not affect test scores in any statistically significant way.

6.7 Practical Implications

Based on the qualitative data collected from the students and a detailed investigation of the area under scrutiny, the researcher compiled the following list of recommendations, which are believed to be applicable to other contexts in which computer-based testing is or is about to be implemented.

- Help the students with Log in processes and ensure everyone has logged in successfully before giving instructions or starting the test. Quite often, for example, after software updates, some computers can take a long time to get started and students panic simply because of that.
- Calm the students down before each CB test sitting and remind them that if anything goes wrong (e.g. the computer freezes or slows down, the test 'disappears', the mouse stops working, etc.), a new attempt will be started for them. This is especially vital when the test is timed and they worry about losing precious time.
- Offer enough practice materials in the same format as the test itself and allow time to train them in operating the technology. Show them how to manipulate the test, as that was the most commonly expressed drawback of the CBT mode in the present study. It is often taken for granted that students, being digital residents, will know how to take a test on a computer, however, their usage of computers in a non-test environment is a completely unrelated matter.
- Provide pieces of paper for note taking during the CB test. A number of students has expressed the need to try out, for example, spelling of

a lexical item, or believed that they can think of an answer better when they write it down by hand.

- Ideally, have technical support at hand or at least a back up plan. Unfortunately, it is not uncommon for some unpredictable technical glitches to resurface during the test sitting.

7 Conclusion

The dissertation presented the first large-scale study regarding Computer-assisted language testing (CALT) in the Czech context. Its aim was to compare two different modes of test administration (computer-based and pencil and paper-based) by analysing quantitative test results as well as qualitative data from more than 350 students in a longitudinal multi-phased study (2014 - 2016) in order to determine whether the usage of computer-based tests at the Department of English and American Studies at Masaryk University is justified.

First, the theoretical framework was thoroughly investigated and a number of fundamental concepts in language testing relevant to the dissertation were introduced and discussed. After situating the study in the historical context of language testing with the help of a categorization proposed by, among others, Spolsky (1977) and Green (2014), four vital test qualities, namely validity, reliability, practicality and washback were examined. These qualities were approached from multiple perspectives put forward by, for example, Bachman (1990), Alderson et al. (1995), James. D. Brown (2005), Weir (2005), H. Douglas Brown (2010) and Carr (2011). Different types of tests, such as Criterion-referenced versus Norm-referenced, Discrete-point versus Integrative, etc., were then detailed in order to describe the main characteristics of the tests used in the empirical part of the dissertation.

The second and major part of the theoretical framework was dedicated to CALT. With regards to Digital Literacy, Bennett and Maton's (2011) criticism of Prensky's (2001) metaphor of digital natives and digital immigrants and White's (2008) preferred terminology of digital residents and digital visitors were commented on. Furthermore, Dudeney's (2013) emphasis on the need to differentiate between being able to use technology for entertainment as opposed to educational purposes was noted. A CALT framework consisting of nine attributes proposed by Suvorov and Hegelheimer (2014) was then explored and direct links were made to the empirical part of the dissertation. The following section detailing the advantages and disadvantages of CALT, presented opinions accompanied with

supporting evidence by renowned scholars, such as James D. Brown (1997), Chapelle and Douglas (2006), Noyes and Garland (2008), and H. Douglas Brown (2010). However, these were mainly viewed from the perspective of test developers and administrators and did not mirror the students' perceptions of the CBT mode very closely as will be shown in the next section. The last subchapter attempted to synopsize comparability studies of the two modes conducted to date, taking into consideration various aspects and individual differences, for example, gender, computer familiarity, and computer attitudes. These studies were then related to the outcomes of the present study in the Findings and Discussion chapter.

Before the central research question is answered, the main findings from the empirical part are depicted. The quantitative data analysis investigated the comparability of the two modes and in later stages of the research explored possible differences with respect to gender and item type. In the Pilot stage (2014), in which all the students took a pencil and paper-based achievement test and then a computer-based one, only the overall test scores were worked with and descriptive statistics was thus made use of. Despite a number of limitations described in 5.1, which could render some of the data questionable, the paired sample t-test showed that the scores did not significantly differ across the two modes. In Study 1 (2015), the research design was altered and the counter-balancing technique was employed to avoid sequencing effects, i.e. half the students took the pencil and paper-based test first and computer-based test second and vice versa. Satisfactory reliability estimates for both versions of the tests were gained and descriptive statistics and statistical models showed only negligible differences between the two modes. Gender differences were not statistically significant. Although an item analysis was available, potential differences in performance relating to item type were not further investigated for the reasons stated in 5.2.

Study 2 (2016) had the most detailed research design with two versions of one achievement test in the two modes administered in the counter-balanced order. The reliability estimates of the tests were very good and the descriptive statistics did not show any discrepancies in test scores. Three statistical models were calculated: the test mode, gender and item type (short answer, multiple choice).

Similarly to Study 1, test scores again slightly differed between the test modes but the difference was only negligible. Gender differences have not been observed, however, the differences related to individual item types were statistically significant in one subtest. It was discovered that students scored better on short answer items in the pencil and paper-based mode than in the computer-based one. Finally, it should be noted that students performed better in the pencil and paper-based mode in all three stages of the research, although the differences were mostly statistically insignificant.

The dissertation is believed to make a valuable contribution to the researched area through the qualitative data analysis, which focused on the advantages and disadvantages of the two modes of test administration as viewed by students. These, for the reasons stated above, differed quite considerably from the accounts provided by the scholars in the theoretical part and therefore garner completely new insights into the research on CALT. Considering they come from more than 350 students, they are certainly worth taking into account for any institutions planning on implementing testing of this type.

Pencil and paper-based test (PPT) advantages were mainly connected to Manipulation, Orientation and Attitudes in all the phases examined. With Pilot, the themes of Familiarity and Physicality were also quite frequent, covering advantages, while Results, Nothing and Handwriting categories fell into the disadvantages of the PPT. The computer-based test (CBT) advantages mirrored the PPT disadvantages to a certain extent and the most frequent theme was Immediate Results. Other aspects related to the possibility of Changing Answers easily and being able to keep track of Time. With Study 1, similar themes were distinguished, however, it is psychologically interesting that although the students appreciated the timer in the CBT mode, the PPT was often regarded as being quicker than the CBT in the Time category. Many other themes, such as Nothing, Attitude, Difficulty or Tradition, reflected the students' feelings and experience. With Study 2, a few new categories were added, mostly related to Technology. There were three main areas that students expressed their likes in regarding the CBT mode, namely Results, Time, and Nothing. One of the most commonly expressed negatives of the CBT here was

categorized under the Technology-related theme and included various difficulties with Log in, the Screen or technical glitches. The chapter cites and acknowledges numerous answers to demonstrate the students' openness and genuine concern.

As for student preferences, which were examined in detail in Study 1 and Study 2, there was a visible shift from the PPT mode preference to the 'no preference' or the CBT mode respectively. In terms of gender, females preferred the PPT mode while males preferred the CBT mode in Study 1. In Study 2, the 'no preference' category was the most popular with females and males, followed by the preference for the CBT mode for both genders. Although there were cases when students preferred the PPT mode and indeed had better scores when writing the test in the PPT mode, the effect of preference on the test results has not been statistically significant except for one subtest in Study 2.

Central RQ:
Is the usage of a computer-based mode of achievement tests justified in the context of Czech tertiary education of the first year English language learners?

The answer to the central research question can therefore be formulated as follows: Based on the research outcomes, the researcher believes that enough evidence has been provided to conclude that the usage of computer-based achievement tests in our context is justified. The minor differences in the test results between the two modes were not statistically significant and the students' feelings and attitudes have become more positive towards the CBT mode.

7.1 Further Research

During the work on the research project, a number of areas of further research were identified. Some of those are directly linked to the limitations depicted in 4.6, 5.2.4 and 5.3.5.

The research design of this large-scale study was rather general as its aim was to compare the two modes of test administration. Having drawn and detailed

comparisons between the two modes, further research could focus on some specific aspects of computer-based tests and investigate, for example, the reasons for the different performance on various item types, the role of note taking, the function of the timer, etc. Furthermore, one of the greatest advantages of the CBT mode is not only immediate results but also instant feedback, however, its potential needs to be exploited more.

Given the size of the present study, the analysis of the qualitative data was mainly descriptive and categorizing. It would be beneficial to gain more profound qualitative data, for example, by conducting interviews with outliers or extreme cases.

Further research could also examine different age groups and see how much age affects both the students' attitudes towards the CBT mode as well as their scores gained. It would be interesting to compare, for example, full time students with combined studies students, who, at least in the context stated, are usually older.

Another suggested area of research would be to investigate teachers' attitudes towards the CBT mode and the possible influence these might have on the students. Human versus computer-based scoring could also be explored further in order to find out whether students' perceptions of teachers being more lenient when marking than the machine (as documented in the qualitative analysis) are justified.

One of the more immediate further research areas is the validation of the new research tool discussed in 5.7. Once the questionnaire has reached the final phase of development, it can be administered to larger populations and gain a lot of information concerning the students' attitudes towards the two modes very efficiently. Based on the analysis of the results, it will be easier to identify extreme cases and work with those in order to help them.

The author of the dissertation believes that the area under scrutiny also needs replication studies. As documented in the theoretical chapter, the equivalence of the two modes cannot be assumed, it has to be manifested (McDonald 2002). However, it is almost impossible to get hold of detailed research procedures of empirical studies conducted in the field and replication is thus not possible. The

researcher has therefore decided to make all the data (anonymized), descriptive statistics, outputs for the statistical models used, etc., accessible on a drive online in order to facilitate research for other researchers. The link can be found in Appendix 0.

7.2 Closing Statement

Unfortunately, testing has traditionally been viewed as a necessary evil rather than a vital part of the teaching process, at least in the Czech Republic. As mentioned in the Introduction, much attention is paid to the act of teaching, while testing is relatively neglected. However, a part of a qualified teacher's job is not only to teach but also to test their students. Furthermore, teachers are expected to design the tests themselves, too often with limited or no background in testing, which is incredibly difficult. Students then, sometimes rightly, complain about the poor quality of the tests and express dissatisfaction with their test results leading to negative washback, which has a detrimental effect on the whole teaching and learning process and results in what seems to be a vicious circle – frustrated teachers, demotivated students and bad test results.

It is time the paradigm shifted and testing came to the fore. The author of the dissertation hopes to have drawn attention to some of the crucial concepts in language testing and, by conducting research in a natural classroom setting and asking the students' opinions regarding the two modes of test administration, connected the theoretical background to the everyday reality of the classroom and demonstrated that the students should be given a voice. The students were very appreciative of this and a number of them, either personally or in their feedback forms, thanked the researcher for showing interest in how they felt about the tests, investigating the topic, and subsequently trying to improve the testing situation in the Department.

This research is a small step on the path to larger scale investigation but once more testing courses are offered for trainee teachers, the importance of testing is recognized and the students are more involved in and better informed about the test development cycle, the myth of testing as a necessary evil will be dispelled.

References

Alderson, J. C. (1986). Computers in language testing. *Computers in English Language Education and Research*, Longman, 99-111.

Alderson, J. C. (2000). Technology in testing: The present and the future. *System, 28(4),* 593-603.

Alderson, J. C. et al. (1995). *Language test construction and evaluation*. New York, USA: Cambridge University Press.

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*. San Francisco, CA: Jossey-Bass.

Anastasi, A. (1988). *Psychological testing* (6th edition). New York: Macmillan.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review, 84(2)*, 191-215.

Barlow, J. P. (1996). *A Declaration of the independence of cyberspace*. Retrieved from http://www.eff.org/Misc/Publications/John_Perry_Barlow/barlow_0296.declaration.txt

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1)*, 1-48. doi:10.18637/jss.v067.i01.

Beckers, J. J., & Schmidt, H. G. (2003). Computer experience and computer anxiety. *Computers in Human Behavior, 19(6)*, 785-797.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology

and assessment. *The Journal of Technology, Learning and Assessment 1(1)*, 1-24.

Bennett, R. E. (2011). Formative assessment: A Critical review. *Assessment in Education: Principles, Policy and Practice, 18(1)*: 5-25.

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6(9),* 1-39.

Bennett, S. & Maton, K. (2011). Intellectual field of faith-based religion: Moving on from the idea of 'digital natives'. In Thomas, M. (ed). *Deconstructing digital natives: Young people, technology and the new literacies*. pp.169-185. London, UK: Routledge.

Bennett, S., Maton, K. and Kervin, L. (2008) The 'digital natives' debate: A critical review of the evidence, *British Journal of Education Technology, 39(5)*, 775-86.

Black, P.J. & Wiliam, D. (2012). Assessment for learning in the classroom. In Gardner, J. (ed.) *Assessment and learning*. 2nd ed. pp. 14-44, London: Sage.

Bloom, B. S. (1968). Learning for Mastery. *Evaluation Comment, 1(2)*, 1-12.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31(1),* 51-60.

Bridgeman, B., Bejar, I. I., & Friedman, D. (1999). Fairness issues in a computer-based architectural licensure examination. *Computers in Human Behavior, 15(3)*, 419-440.

Brookhart, S. (2008). *How to give effective feedback to your students*. Alexandria, VA: Association for Supervision and Curriculum Development.

Broos, A. (2005). Gender and information and communication technologies (ICT) anxiety: Male self-assurance and female hesitation. *CyberPsychology & Behavior, 8(1)*, 21-31.

Brosnan, M. J. (1998). The impact of psychological gender, gender-related

perceptions, significant others, and the introducer of technology upon computer anxiety in students. *Journal of Educational Computing Research, 18(1)*, 63-78.

Brown, H. D. (2007). *Principles of language learning and teaching*. White Plains: Pearson Longman.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: principles and classroom practices*. Upper Saddle River, N.J: Pearson Education.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1(1)*, 44-59.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* Upper Saddle River, N.J.: Prentice Hall Regents.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.), pp.367-407. London: Collier Macmillan.

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed.), pp. 529–38. Oxford, England: Oxford University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In Richards, J. C., & Schmidt, R. W. (Eds.), *Language and Communication*, pp. 2-27. London: Longman.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1,* pp. 1-47.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44(1)*, 57-66.

Carr, N. T. (2006). Computer-based testing: Prospects for innovative assessment. *Calling on CALL: From theory and research to new directions in foreign language teaching*, 289-312.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Carroll, J. B. and S. Sapon. (1959). *Modern language aptitude test (MLAT)*. New York, NY: The Psychological Corporation.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-72.

Chapelle, C. A. (2010). *Technology in language testing* [video]. Retrieved from http://languagetesting.info/video/main.html

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge, England: Cambridge University Press.

Charmaz, K. (2005). Grounded Theory in the 21st Century: Applications for advancing social justice studies. In Denzin, N. K., & Lincoln, Y. S. *The SAGE handbook of qualitative research*. Thousand Oaks: Sage Publications.

Chien, T. C. (2008). *Factors influencing computer anxiety and its impact on e-learning effectiveness*: A Review of Literature. Online Submission. http://files.eric.ed.gov/fulltext/ED501623.pdf

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: Massachusetts Institute of Technology Press.

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

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states*. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved from http://files.eric.ed.gov/fulltext/ED474867.pdf

Cohen, S. (1972). *Folk devils and moral panics*. London: McGibbon & Kee.

Creswell, J. W., & Clark, V. L.P. (2011). *Designing and conducting mixed methods research*. Thousand Oaks: Sage.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed*

methods approaches. (4th ed.) Los Angeles: Sage.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3)*, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers, 34(2)*, 177-180.

Davey, T. (2011). *Practical considerations in computer-based testing*. Princeton, NJ: Educational Testing Service.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

DiLalla, D. L. (1996). Computerized administration of the multidimensional personality questionnaire. *Assessment, 3(4)*, 365-374.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Douglas, D. (2000). *Assessing languages for specific purposes*. New York, NY: Cambridge University Press.

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27*, 115-132. Retrieved from http://hstrik.ruhosting.nl/wordpress/wp-content/uploads/2013/03/Douglas-Hegelheimer-ARoAL-2007.pdf

Draper, S. W. (2009). Catalytic assessment: understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology, 40(2)*, 285-293.

Dudeney, G. Hockly, N & Pegrum, M. (2013). *Digital literacies.* Malaysia: Pearson.

Dunkel, P.A.: 1999, Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning and Technology 2(2)*, 77–93.

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quaterly, 16(1)*.

Farhady, H. (2006). *Twenty-five years of living with applied linguistics: Collection of articles*. Iran: Rahnama Press.

Farhady, H., and Shakery, S. (2000). Number of options and economy of multiple-choice tests. *Foreign Language Teaching Journal, 14 (57).*

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal, 53(4)*, 289–299.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Gardner, J.(ed.), (2012). *Assessment and learning*. Los Angeles, Calif: Sage.

Glowacki, M. L., McFadden, A. C., & Price, B. J. (1995). Developing computerized tests for classroom teachers: A pilot study. Paper presented at the *Annual Meeting of the Mid-South Educational Research Asssociation*, Biloxi, MS. (ERIC Document Reproduction Service No. ED391471). Retrieved from Education Resources Information Center [ERIC] Web site: http://www.eric.ed.gov

Gos, M. W. (1996). Computer anxiety and computer experience: a new look at an old relationship. *The Clearing House, 69(5)*, 271–276.

Green, A. (2014). *Exploring language assessment and testing*: Language in Action. Abingdon, Oxon: Routledge.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12(2)*, 109-113.

Gruba, P. A. (2000). *The role of digital video media in second language listening comprehension*. PhD thesis, Department of Linguistics and Applied Linguistics, University of Melbourne.

Hargis, C. H. (2003). *Grades and grading practices: Obstacles to improving education and to helping at-risk students*. (2nd ed.). Springfield, IL: Charles C. Thomas Publisher, Ltd.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: does the medium in which assessment questions are presented affect children's performance in mathematics?. *Educational Research, 46(1)*, 29-42.

Harlen, W. (2012). On the relationship between assessment for formative and

summative purposes. In Gardner, J. (2012). *Assessment and learning*. Los Angeles, Calif: Sage.

Hayward, L. (2012). Assessment and learning: The learner's perspective. In Gardner, J. *Assessment and learning*. Los Angeles, Calif: Sage.

Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford, UK: Oxford University Press.

Henning, G. (1987). *A Guide to Language Testing*. Cambridge, MA.: Newbury House.

Howard, G. S. (1986). *Computer anxiety and management use of microcomputers*. Ann Arbor: UMI Research press.

Hughes, A. (1989, 2013). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, A., Porter, D. & Weir, C. J. (1988). *Validating the ELTS test: A critical review*. Cambridge: The British Council and the University of Cambridge Local Examination Syndicate.

Hymes, D. H. (1972). On Communicative competence. In Pride, J. B., & Holmes, J. (eds). *Sociolinguistics*. pp.269-293. Harmondsworth, UK: Penguin Books.

Ingram, E. (1977). Basic concepts in testing. In J.P.B. Allen & A. Davies (eds.), *Testing and experimental methods*. pp.195-216. Oxford: Oxford University Press.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistic*s, 25, 228-242.

Johnson, B., & Christensen, L. B. (2008). *Educational research: quantitative, qualitative, and mixed approaches* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Joiner, R., Gavin, J., Brosnan, M., Cromby, J., Gregory, H., Guiller, J., Maras, P., Moon, A. (2013). Comparing first and second generation digital natives' Internet use, Internet anxiety, and Internet identification. *Cyberpsychology, Behavior and Social Networking, 16 (7)*, 549-52.

Kay, R. H. (1993). An exploration of theoretical and practical foundations for assessing attitudes towards computers: the computer attitude measure (CAM). *Computers in Human Behavior, 9*, 371–386.

Kernan, M. C., & Howard, G. S. (1990). Computer anxiety and computer attitudes: An investigation of construct and predictive validity issues. *Educational and Psychological Measurement, 50(3)*, 681-690.

Kozulin, A., & Vygotskii, L. S. (1986). *Thought and language*. Cambridge: MIT Press.

Kuznetsova, A., Brockhoff, P.R., & Christensen, R.H.B. (2016). *lmerTest: Tests in linear mixed effects models*. R package version 2.0-33. Retrieved from https://CRAN.R-project.org/package=lmerTest

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lantolf, J., & Poehner, M. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics, 1 (1)*, 49-72.

Larson, J. W. (1987). Computerized adaptive language testing: A Spanish placement exam. In K. M. Baily, T. L. Dale, & R. T. Clifford (Eds.), *Language testing research.* pp. 1-10. Monterey, CA: Defense Language Institute.

Levine, T., & Donitsa-Schmidt, S. (1998). Computer use, confidence, attitudes, and knowledge: A causal analysis. *Computers in Human Behavior, 14(1)*, 125-146.

Lightstone, K., & Smith, S. M. (2009). Student choice between computer and traditional paper-and-pencil university tests: what predicts preference and performance?. *Revue internationale des technologies en pédagogie universitaire/International Journal of Technologies in Higher Education, 6(1)*, 30-45.

Lyons, J. (1996). On communicative competence and performance. In Brown, G., Malmkjær, K., & Williams, J.: *Performance and competence in second language acquisition*. Cambridge: Cambridge University Press.

M. Mohler and R. Mihalcea. (2009). *Text-to-text semantic similarity for automatic short answer grading*. In Proceedings of the European Association for Computational Linguistics (EACL 2009), Athens, Greece.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach.* Thousand Oaks, Calif: SAGE Publications.

McAllister, D., & Guidice, R. M. (2012). This is only a test: A machine-graded

improvement to the multiple-choice and true-false examination. *Teaching in Higher Education, 17(2)*, 193-207.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education, 39(3)*, 299-312.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114(3)*, 449-458.

Messick, S. (1996). *Validity and washback in language testing*. Princeton, N.J: Educational Testing Service.

Morrow, K. (1979). Communicative language testing: revolution of evolution? In: Brumfit, C.K., Johnson, K. (Eds.), *The Communicative Approach to Language Teaching*. pp. 143-159, Oxford University Press, Oxford.

Morrow, K. (2012). Comunicative language testing. In Coombe, C. A., In Davidson, P., In O'Sullivan, B., & In Stoynoff, S. *The Cambridge guide to second language assessment*. pp.140-146, NY: Cambridge University Press.

Nepivodová, L. (2006). On communicative language competence, validity and different modes of administration. MA thesis, Department of English and Americal Studies, Masaryk University.

Nickell, G. S., & Pinto, J. N. (1986). The computer attitude scale. *Computers in human behavior*, *2(4)*, 301-306.

Noijons, J. (1994). Testing computer assisted language testing: Towards a checklist for CALT. *Calico Journal, 12(1)*, 37-58.

Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent?. *Ergonomics, 51(9)*, 1352-1375.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal, 93,* 836–47.

Ogles, B. M., France, C. R., Lunnen, K. M., Bell, T., & Goldfarb, M. (1998). Computerized depression screening and awareness. *Community mental health journal, 34(1)*, 27-38.

Oller, J. W. (1979). *Language tests at school*. Harlow, UK: Longman.

Oller, J.W. (ed.), (1983). *Issues in language testing research*. Rowley, Mass.: Newbury House.

Paek, P. (2005). Recent trends in comparability studies. *Pearson Educational Measurement*. Retrieved from https://www.researchgate.net/profile/Pamela_Paek/publication/24502391 1_Recent_Trends_in_Comparability_Studies_Using_testing_and_assessment_t o_promote_learning/links/00b7d51d5c29b537b5000000.pdf

Palfrey, J. G. & Gasser, U. (2008). Reclaiming an awkward term: What we might learn from 'digital natives'. In Thomas, M. (ed). Deconstructing digital natives: Young people, technology and the new literacies. pp.186-204, London, UK: Routledge.

Parshall, C. G., & Harmes, J. C. (2014). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology, 10(1)*, 1-20.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2009). Innovative items for computerized testing. In *Elements of adaptive testing* (pp. 215-230). Springer New York.

Phelps, R. P. (2011). Teach to the test?. *The Wilson Quarterly*, *35(4)*, 38-42.

Plakans, L. (2009). *Integrated assessment* [video]. Retrieved from http://languagetesting.info/video/main.html

Poehner, M. E. (2008). *Dynamic assessment: A vygotskian approach to understanding and promoting L2 development*. Berlin: Springer.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research, 32(2)*, 153-166.

Powers, D. E. (1999). *Test anxiety and test performance: comparing paper-based*

and computer-adaptive versions of the GRE general test (RR-99-15). Princeton, NJ: Educational Testing Service.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon, 9(5)*, 1-6. Retrieved from http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf.

Prensky, M. (2011). Digital wisdom and homo sapiens digital. In Thomas, M. (ed). *Deconstructing digital natives: Young people, technology and the new literacies*. pp.15-29, London, UK: Routledge.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Rabinowitz, S., & Brandt, T. (2001). *Computer-based assessment: Can it deliver on its promise?* Knowledge Brief. Retreived from http://files.eric.ed.gov/fulltext/ED462447.pdf

Reece, M. J., & Gable, R. K. (1982). The development and validation of a measure of general attitudes toward computers. *Educational and Psychological Measurement, 42(3)*, 913-916.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement: issues & practice. 24(2)*, 3-13.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5(2), 84–94*.

Russell, M. (1999). *Testing on computers: a follow-up study comparing performance on computer and on paper*. Education Policy Analysis Archives (online). Retrieved from: <http://epaa.asu.edu/epaa/v7n20.html> (3 September 2015).

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper. *Education policy analysis archives, 5(3)*.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, and

M. Scriven (Eds.), *Perspectives of curriculum evaluation, Volume I*, pp. 39-83. Chicago, IL: Rand McNally.

Šeďová, K., Švaříček, R., & Šalamounová, Z. (2012). *Komunikace ve školní třídě*. Praha: Portál.

Selwyn, N. (1997). Students' attitudes toward computers: Validation of a computer attitude scale for 16–19 education. *Computers & Education, 28(1)*, 35-41.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions.* New York, NY: Routledge.

Silverman, D. (2006). *Interpreting qualitative data: Methods for analyzing talk, text and interaction.* London: Sage Publications.

Smith, B., & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior, 23(3)*, 1481-1498.

Soukup, P., & Kocvarová, I. (2016). Velikost a reprezentativita výberového souboru v kvantitativne orientovaném pedagogickém výzkumu 1. *Pedagogická orientace, 26(3)*, 512-536.

Spolsky, B. (1977). Language Testing: art or science, in Nickel, G. (ed.) *Proceedings of the Fourth International Congress of Applied Linguistics*, pp. 7-28. Stuttgart: Hochschulverlag.

Stevenson, D. K. (1985). Authenticity, validity and a tea party. *Language Testing, 2, 1*, 41-47.

Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Research Report. Iowa City, IA: Pearson Education.

Stricker, L. J., Wilder, G. Z., & Rock, D. A. (2004). Attitudes about the computer-based test of English as a foreign language. *Computers in Human Behavior, 20(1)*, 37-54.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford, OX, UK: Blackwell Publishers.

Suvorov, R., & Hegelheimer, V. (2014). *Computer-assisted language testing*. The

companion to language assessment. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla083/full

Švaříček, R., & Šeďová, K. et al. (2007). *Kvalitativní výzkum v pedagogických vědách*. Praha: Portal, s.r.o.

Tapscott, D. (1998). *Growing up digital. The rise of the net generation*. New York: McGraw Hill.

Tapscott, D. (2009). *Grown up digital: How the net generation is changing your world*. New York: McGraw-Hill.

Tarone, E. (1988). *Variation in interlanguage*. London: Edward Arnold.

Tavakol, M., & Dennick, R. (2011). *Making sense of Cronbach's alpha*, 53–55. Retrieved from http://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (Research Reports 61). Princeton, NJ: Educational Testing Service.

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49(2)*, 219–274.

Valette, R. (1967) *Modern language testing: A Handbook*. New York, NY: Harcourt.

Van, De, Vijver, F. J. R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79(6)*, 852–859.

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: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68(1)*, 5-24.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke: Palgrave Macmillan.

Weir, C., Yan, J., O'Sullivan, B., & Bax, S. (2007). Does the computer make a difference?: The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: effects and

impact. *International English Language Testing System (IELTS) Research Reports 2007: Volume 7, 1*.

White, D.S. (2008). *Not 'natives' and 'immigrants' but 'visitors' and 'residents*. TALL blog. Retrieved from http://tallblog.conted.ox.ac.uk/index.php/2008/07/23/not-natives-immigrants-but-visitors-residents/

White, D.S. and Le Cornu, A. (2011). Visitors and residents: A new typology for online engagement. *First Monday, 16(9)*, Retrieved from http://firstmonday.org/article/view/3171/3049.

Whitley, B. E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior, 13(1)*, 1-22.

Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior, 22(5)*, 791-800.

Winke, P., & Fei, F. (2008). Computer-assisted language assessment. In N. Van Deusen- Scholl & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 4, pp. 353–64). New York, NY: Springer.

Young, B. J. (2000). Gender differences in student attitudes toward computers. *Journal of Research on Computing in Education, 33(2)*, 204-216.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education, 29(4)*, 423-438.

Žitný, P., Halama, P., Jelínek, M. & Květon, P. (2012). Validity of cognitive ability tests – comparison of computerized adaptive testing with paper-and-pencil and computer-based forms of administrations. *Studia Psychologica, vol. 54, no. 3*, 181-194.