TALLINNA ÜLIKOOL
HUMANITAARTEADUSTE DISSERTATSIOONID

TALLINN UNIVERSITY
DISSERTATIONS ON HUMANITIES

22

TALLINNA ÜLIKOOL

ENE ALAS

THE ENGLISH LANGUAGE NATIONAL EXAMINATION VALIDITY DEFINED BY ITS ORAL PROFICIENCY INTERVIEW INTERLOCUTOR BEHAVIOUR

TALLINN 2010

TALLINNA ÜLIKOOL
HUMANITAARTEADUSTE DISSERTATSIOONID

TALLINN UNIVERSITY
DISSERTATIONS ON HUMANITIES

22

ENE ALAS

THE ENGLISH LANGUAGE NATIONAL EXAMINATION VALIDITY DEFINED BY ITS ORAL PROFICIENCY INTERVIEW INTERLOCUTOR BEHAVIOUR

Institute of Germanic and Romance Languages and Cultures

The dissertation is accepted for the commencement of the degree of Doctor of Philosophy (linguistics) by the Doctoral Committee of Humanities of Tallinn University on April 19, 2010.

Supervisor: Suliko Liiv, Ph.D., Tallinn University, Professor
Opponents: Mirja Tarnanen, Ph. D., senior researcher at the Centre for Applied Language Studies at the University of Jyväskylä
Tuuli Oder, Ph. D., head of Tallinn University Language Centre
The academic disputation of the dissertation will be held on June 9, 2010 at 12 o'clock, at Tallinn University, Narva mnt. 29, lecture hall S-403.

ISSN 1736-3621 (trükis)
ISSN 1736-5031 (dissertatsioon, online PDF)
ISSN 1736-3667 (analüütiline ülevaade, online PDF)
ISBN 978-9949-463-03-9 (trükis)
ISBN 978-9949-463-04-6 (dissertatsioon, online PDF)
ISBN 978-9949-463-05-3 (analüütiline ülevaade, online PDF)

Tallinn University Press
Narva mnt. 25
10120 Tallinn
www.tlu.ee/kirjastus

THE ENGLISH LANGUAGE NATIONAL EXAMINATION VALIDITY DEFINED BY ITS ORAL PROFICIENCY INTERVIEW INTERLOCUTOR BEHAVIOUR

Abstract

The current doctoral dissertation is a study of the inception and advancement of the English language national examination in Estonia. The main goal of research is to trace the changes that have been introduced in the examination system, reflecting the development of the understanding of the English language proficiency construct in Estonia. More specifically, the dissertation concentrates on the impact of one variable – the oral proficiency interviewer – on the consequent proficiency evaluation validity within the framework of the examination.

The dissertation opens with an introduction that defines research goals and hypotheses, research methods, materials and the structure of the dissertation. Chapter one presents the state of the art with regard to the current understanding of language proficiency, oral proficiency interview as an assessment tool and the role of the interviewer behaviour in the process of language proficiency evaluation. Chapter two discusses the English language national examination development during the period of 1995 to 2008, assessing the validity of the changes made in terms of examination construct, marking procedures and perennial examination results. Chapter three presents the results of a questionnaire study conducted among the oral proficiency interviewers that yielded an overview of the interviewer's own perception of their role in the proficiency evaluation process. Statistical and cluster analysis indicated that the interviewers displayed patterns in their perception of the role they had during the interview. Chapter four recounts the outcomes of a qualitative study of the interviewer behaviour and language during the 2008 English language national examination oral proficiency interviews. The study focuses on the degree of adherence to the interviewer scripts, discusses the nature of the deviations with regard to the emerging patterns in the actual interviewer behaviour and searches for links between the interviewer behaviour and the interviewer gender and school type. Conclusion will summarise the current research results in terms of the research hypotheses, outline the implications of the study and propose directions for further research.

Keywords: language testing, construct of language proficiency, validity, reliability, language test validation, test design, test retrofit, construct-irrelevant variance, oral proficiency interview, interlocutor behaviour, interlocutor effect, interlocutor training, interlocutor variability, interview scripts, accommodation, rating scale.

SUULISE KEELEPÄDEVUSTESTI EKSAMINEERIJA INGLISE KEELE RIIGIEKSAMI VALIIDSUSE MÄÄRAJANA

Resümee

Käesolev doktoritöö uurib inglise keele riigieksami tekkelugu ja selle arendust Eestis. Töö peaeesmärk on uurida , milliseid muudatusi on inglise keele riigieksamitöös kui keelepädevustestis tehtud alates selle sisseviimisest proovieksamina 1995 aastal. Need muudatused peegeldavad seda, kuidas on muutunud ja arenenud inglise keele riigieksami koostajate jaoks keelepädevuse konstrukti mõiste senise keeletestimise praktika jooksul. Dissertatsiooni peatähelepanu on koondatud ühe olulise eksamitulemuse määratleja – suulise eksami intervjueerija – tegevusele ja sellest tulenevale eksamitulemuse valiidsusele inglise keele riigieksamil.

Doktoritöö algab sissejuhatusega, kus määratletakse uurimistöö eesmärgid ja hüpoteesid, uurimismeetodid, materjal ja dissertatsiooni struktuur. Esimene peatükk esitab teoreetilise ülevaate uurimisvaldkonna praegustest seisukohtadest, mis puudutavad keelepädevuse mõistet, suulise keelepädevuse intervjuud kui hindamismeetodit ning suulise eksami intervjueerija rolli suulise keelepädevuse hindamise protsessis. Teine peatükk käsitleb inglise keele riigieksami arendust aastatel 1995 kuni 2008, hinnates nende aastate jooksul tehtud muudatuste valiidsust eksami konstrukti, hindamisprotseduuri ja iga-aastaste eksamitulemuste seisukohalt. Kolmas peatükk esitab ja analüüsib uurimust, mis tehti küsitlusena inglise keele riigieksami eksamineerijate hulgas ning andis ülevaate eksamineerijate arusaamisest enda kui eksamineerija rollist keelepädevuse hindamise protsessis. Nii statistiline kui ka klasteranalüüs näitasid, et selle põhjal, millisena eksmineerijad oma rolli hindamisprotsessis näevad, võib nad jaotada selgelt eristuvatesse rühmadesse. Neljas peatükk vaatleb kvalitatiivse uurimuse tulemusi, mis saadi 2008. aasta riigieksami suulise eksami intervjuude jooksul jäädvustatud intervjueerijate tegevust analüüsides. Uurimus analüüsib seda, millisel määral pidasid intervjueerijad kinni nõutud käsikirjast, millised olid intervjueerijate tehtud muudatused ja kuivõrd oli tehtud muudatustes süsteemsust. Samuti uuritakse, millisel määral sõltub intervjueerija käitumine intervjueerija soost ja kooli tüübist. Doktoritöö kokkuvõte summeerib uurimistöö tulemused esitatud hüpoteeside valguses, analüüsib nende tähendust ja mõju riigieksami arendusprotsessis ning kirjeldab mõningaid võimalikke arengusuundi edasiseks testimisalaseks teadustööks Eestis.

Võtmesõnad: keeletestimine, keelepädevuse konstrukt, valiidsus, usaldusväärsus, keeletesti valideerimine, testi koostamine, testi uuendamine, konstruktist sõltumatu muutuja, suulise keelepädevuse intervjuu, eksamineerija käitumine, eksamineerija mõju, eksamineerijate koolitus, intervjueerija variatiivsus, intervjuu käsikiri, akommodatsioon/ kohandumus, hindamisskaala.

CONTENTS

ACKNOWLEDGEMENTS

This dissertation would never have been completed without conceptual guidance and inspiration from my supervisor, professor Suliko Liiv, director of Tallinn University Institute of Germanic and Romance Languages and Cultures.

A special thanks to the colleagues at National Examination and Qualification Centre for giving me access to their database, archives and library.

I am hugely indebted to my reviewers Mirja Tarnanen, senior researcher at the Centre for Applied Language Studies at the University of Jyväskylä and Tuuli Oder, head of Tallinn University Language Centre for undertaking the enormous task of evaluating my work.

Last, but certainly not least, a heartfelt thank-you to my colleagues at Tallinn University and my family for giving me the support to complete this dissertation.

LIST OF PUBLICATIONS

Alas, Ene. 2010. Interviewer Variability in Oral Proficiency Interviews. In Norquist, R. (ed.) *Crossing Boundaries: Studies in English Language, Literature and Culture in a Global Environment*. (9–35). Peter Lang Internationaler Verlag der Wissenschaften.

Liiv, Suliko, Alas, Ene. 2009. Evaluation de l'anglais dans un examen national: résultats et problèmes. in *Synergies Pays riverains de la Baltique*, numéro 6 "Problématiques culturelles dans l'enseignement-apprentissage des langues-cultures, mondialisation et individualisation: approche interdisciplinaire", Gerflint, France, (241–247).

Alas, Ene; Liiv, Suliko. 2009. Constraints of Measuring Language Proficiency in Estonia: The National Examination in the English Language. H.Metslang, M.Langemets, M.-M. Sepper (Toim.). Eesti Rakenduslingvistika Ühingu aastaraamat 5, Estonian Papers in Applied Linguistics 5 (19–32). Tallinn: Eesti Keele Sihtasutus

Alas, Ene, Roosmaa, Ester. 2008. Oral Examination Script Specimen 2008. http://www.ekk.edu.ee

Alas, Ene. 2008. Oral Examination Introductory Stage Specimen. http://www.ekk.edu.ee

Alas, Ene, Roosmaa, Ester. 2008. Interviewers' and Assessors' Procedures for the Nationa Examination 2008. http://www.ekk.edu.ee

Alas, Ene. 2008. Guidelines for the Oral Part of the National Examination 2008. http://www.ekk.edu.ee

Alas, Ene. 2008. National Examination Marking Scale for Speaking. http://www.ekk.edu.ee

Alas, Ene. 2007. A Marking Scale for Letters, A Marking Scale for Essays and Reports. National Examination 2007. http://www.ekk.edu.ee

Alas, Ene; Roosmaa, Ester. 2007. Inglise keele riigieksam: uuendused rääkimisoskuse testimises. Õpetajate leht 44.

Alas, Ene. 2007. Developing the National Examination in the English Language. Open!, 32, 2–5.

Alas, Ene. 2007. ALTE eksamite hea tava põhimõtted : redigeeritud variant, oktoober 2001. 17 lk. Tõlkija.

Alas, Ene, Roosmaa, Ester. 2006. Muutusi inglise keele riigieksami kirjaliku osa hindamisjuhendis. Õpetajate leht 44.

Alas, Ene. 2005. Interdepartmental Interpretation of Academic Essays. Jane Honka, Nancy Aalto, Elisabeth Heap-Talvela, Felicity Kjisik, Joan Norlund (ed.). Celebrating the Second 10 Workshops the Communication Skills Workshop: Contexts, Signposts, Words (73–80). Helsinki: Watam Press.

Alas, Ene. 2004. Assessment of Academic Writing. Neil Murray, Tony Thorne (ed.). Multicultural Perspectives on English Language and Literature (15–24). Tallinn, London: TPÜ Kirjastus.

Alas, Ene. 1999. ESL/EFL perspective of the assessment of students' academic writing. Nordquist, Richard (ed.). International Perspectives on English and American Language and Literature (7–20). Tallinn: TPÜ Kirjastus.

Alas, Ene. 1994. Validation of Formal Criteria for the Assessment of Writing for Academic Purposes, MA Dissertation, 200 pp, Reading University, Great Britain.

LIST OF ABBREVIATIONS

B1 – marker for the first level of the independent user on the global scale of the Common European Reference for Languages

B2 – marker for the second level of the independent user on the global scale of the Common European Reference for Languages

CEFR – Common European Framework of Reference for Languages

CPE – Certificate of Proficiency in English (a Cambridge exam)

E followed by a number, e.g. E1 – an interview/ interviewer in an Estonian medium school

FCE – First Certificate in English (a Cambridge exam)

IELTS – International English Language Testing System

MC – multiple choice questions/ tasks

NEQC – National Examination and Qualification Centre

OPI – oral proficiency interview

Q followed by a number, e.g. Q1 – a question in the interviewer questionnaire

R followed by a number, e.g. R 1 – an interview/ interviewer in a Russian medium school

TFN – true/false/no information questions/tasks

TOEFL – Test of English for the speakers of Foreign Languages

INTRODUCTION

RESEARCH GOALS AND HYPOTHESES

The current doctoral dissertation has been inspired by the testing and evaluation practices that started to find their way into Estonian gymnasiums and upper secondary schools with the advent of national examinations, which took place starting from the middle of the 1990s along with a number of other Eastern European countries (Latvia, Lithuania, Hungary, Poland, etc.). The national examination in the English language, which the current thesis has chosen at its focus, has been officially operational since 1997 and thus has a history of over a decade here already. Although post hoc statistical item analysis is annually conducted with regard to the quality of the English language national examination paper, and the rater variance is being controlled within the writing section of the examination by providing multiple raters, if necessary, there has been virtually no research into how the national examination in the English language has evolved over the years and what the main trends of development have been. Changes have been introduced into virtually every section of the examination, reflecting the evolution of understanding of the construct of the different skills being tested. Yet there is virtually no analysis of the nature of the changes that have been introduced over the years. Recording and discussing the development of the English language national examination framework is the first goal of the current doctoral dissertation.

The other main direction of the current research has been promoted by the most problematic section of the national examination in the English language in Estonia – the oral proficiency interview (OPI). The OPI has long been a favourite means of measuring the candidates' oral proficiency in the world and has also been adopted here to evaluate the candidates' level of speaking ability. The OPI has stood the test of time internationally and many of its features have been well researched. There is, however, an aspect of OPI that has emerged in the respective research literature fairly recently and deserves a closer investigation because it affects the reliability and consequently also the validity of any proficiency examination which includes an OPI as its section, and that is the interviewer behaviour during the interviews. The doctoral thesis at hand will attempt to narrow that gap.

Subsequent chapters of the current thesis will investigate the national examination in the English language in Estonia in the light of current theoretical considerations concerning proficiency examination development and oral proficiency interview interlocutor behaviour. More particularly, it will study the development of the national examination in the English language in Estonia into the language proficiency instrument that it is today and attempt at evaluating if valid inferences about students' language proficiency can be drawn from its results. It will also study the characteristic features of interviewer behaviour and language as they

emerge during the oral interview of the examination. The research goals can be summarised in the following hypotheses:

Hypothesis 1. The current national examination in the English language will allow a valid evaluation of students' language proficiency with the speaking test containing the greatest validity threat.

Hypothesis 2. Interviewers conducting the OPI during the national examination in the English language will vary in their understanding of the expectations to their own and student behaviour during the oral proficiency interview.

Hypothesis 3. Interviewer language and behaviour during the oral proficiency interviews will display a high degree of adherence to the interviewer scripts provided for the speaking test.

Hypothesis 4. Deviations from the provided interviewer scripts will display patterns.

METHODS OF RESEARCH

Research methods have been prompted by the research hypotheses posed, so that they would allow a comprehensive treatment of the data gathered, a clear presentation of results as well as viable implications for further testing practices and research.

Hypothesis 1 will be investigated relying on the descriptive methods of research by first looking at the process of the English language national examination genesis and the subsequent examination retrofit. The examination's suitability as a proficiency assessment tool will be evaluated relying on the analysis of the examination framework and individual examination papers from 1995 to 2008 and the alterations made in it. For the same purpose, the examination results will be compared and contrasted over the same period of time to establish the level of consistency and hence validity of the exam.

Hypothesis 2 will be studied with the help of first discussing the new framework proposed for the speaking section of the English language national examination as of 2008. This will be followed by a questionnaire study carried out among the OPI interviewers. Eighty-one participants' interview responses will be analysed for the estimation of their perception of their role and behaviour during the interview. Spearman rank correlation and cluster analyses will be utilised to discover the presence of behavioural patterns (procedural and linguistic) during the interview.

Hypotheses 3 and 4 will be studied resorting predominantly to qualitative data analysis methods, which have been adapted for the purposes of the present study from those proposed by A. Lazarton and A. Brown in their respective studies (cf. Lazarton 2002 and Brown 2005) of interviewer behaviour and language during the OPI. Comparative methods as well as conversation analysis will be applied to the recordings of 50 national examination oral proficiency interviews with the aim of discovering the degree of adherence to the interview script and the presence of patterns among

the deviant behaviour. SPSS –16 data analysis system will be used to quantitatively corroborate particular qualitative analysis findings.

The methodological choices have been made proceeding from the principle of triangulation – 'the attempt to understand some aspect of human behaviour by studying it from more than one standpoint, often making use of both quantitative and qualitative data in doing so (Brown and Rodgers 2002:243). Viewing the data from at least two viewpoints 'will maximise the possibility of getting credible findings by cross-validating those findings' (ibid). Thus the national examination validity quest is substantiated by data triangulation paired with methodological triangulation (the current dissertation draws from the information obtained by studying the materials documenting the national examination development, examiner questionnaires and interview transcript analyses).

The reasons for resorting to the above-mentioned methods are manifold. First of all, the materials analysis allows a very thorough overview of the practical work done in developing national English language proficiency tests. The national examination in the English language, in some sense, has a very short history, its inception dating back to 1993; thus looking at all the examinations is not an insurmountable problem. Studying all the pilot exams and all official national examination papers would yield a more comprehensive view of the evolution of the language proficiency construct and its measurement principles within the English language national examination, than by just investigating, say, every other/ third/ fifth year if the period had been longer. Discussion of the materials allows us to initiate the national examination validity evaluation process with the aim of pinpointing the problem areas with a view of possible further investigation of that area.

The questionnaire study aims at discovering the interviewers' perception of the OPI, their perception of the process of it and their view of the students' behaviour and of their own behaviour during the interview. The questionnaire is used to discover to what extent the interviewers are aware of their many roles and tasks during the conduct of the interview. By looking at the interviewer perception we can perhaps decide if there is a conscious attempt at uniformity during oral proficiency testing.

As the questionnaire allows a subjective view of the interviewers' disposition, the interviewers' perception of their behaviour during the OPI, and has the danger of providing a slightly scewed picture of the interviewer behaviour, another method of research – interview transcript analysis – was deemed necessary, it was thought necessary to verify if what the interviewers claimed to be doing during the interview was actually substantiated by their actual performance. Interview transcript analysis is envisaged to juxtapose the interviewers' perception with their actual behaviour. Interview transcript analysis is expected to produce behavioural patterns, which might be marked for gender or the cultural background of the interviewer. The discovery of patterns, in its turn, would help to develop more informed training programmes for the interviewers or change the interview pattern altogether if necessary.

RESEARCH MATERIALS

The current dissertation material has been drawn from the archived documents concerning the national examinations of 1995–2008 in the National Examination and Qualification Centre, a questionnaire carried out among 81 interviewers who were involved in the 2008 national examination in the English language, and recordings of 50 national examination interviews of 2008. The materials used are summarised in the table below.

Table 1. Research materials informing the current dissertation.

Hypothesis	Materials
1	• Documentation regarding national examination legislation (1995–2008). • The English language national examination papers (1995–2008). • The National Curriculum. • Year 12 examination specifications. (Handbooks) • Public responses to the national examination development in the press. (Newspaper articles). • Examination statistics (1995–2008).
2	• The English language national examination OPI framework for interviewers (3 scripts). • Guidelines for OPI interviewers. • 81 response forms of the OPI interviewer questionnaire containing 40 statements and 4 open questions. • Spearman rank correlation data • Cluster analysis data
3 and 4	• 50 national examination interviews in the form of - 183 pages of interview transcripts - 10 hours 32 minutes and 27 seconds of interview recordings

Preliminary findings based on the database described above have been published in three articles: Alas and Liiv (2009), Liiv and Alas (2009) and Alas (2010).

STRUCTURE OF THE DISSERTATION

The current dissertation will proceed to discuss the current national examination research and development in the English language in four sections. Chapter one will consider advances in testing and evaluation as outlined in the British and American sources of research literature, as it is those sources that predominantly inform the respective test development in Estonia at the moment. The main foci of scrutiny will be the underlying theoretical considerations while making test development decisions, and the fundamental interviewer characteristics as they emerge in research literature today.

Chapter two takes a chronological view of the English language national test development process in Estonia and attempts to discuss the changes made over time to create an instrument for valid proficiency evaluation. It will discuss the national examination preparation period with the base-line study that paved the way to the 1995 and 1996 pilot examinations. It will then compare and contrast all the subsequent

national examination papers to trace the process of test retrofit as the knowledge regarding a valid proficiency evaluation instrument accumulated among the members of the examination development team. Validity of the examination will be assessed in terms of the examination construct, marking procedures and perennial examination results.

Chapter three discusses a questionnaire study conducted among Estonian teachers of English (non-native speakers of English), who act as interviewers or assessors during the speaking section of the national examination in the English language, to investigate the interviewers' perception of their role and conduct during the interview. The chapter will first discuss the new framework introduced during the English language national examination OPI in 2008 in terms of the requirements it poses for the interviewer. It will then outline the questionnaire study and discuss the data provided by the respondents. Research will attempt to discover if groups emerge within the responding set, and if so, what the characteristic features of each group would be. Group characteristics will hopefully have implications for interviewer training and general testing practices and future examination development.

Chapter four is an account of a qualitative study of interviewer behaviour and language, based on the recordings of the speaking section during the national examination in the English language in Estonia in 2008. The chapter will attempt a meticulous analysis of the interviews with regard to the degree of adherence to the interviewer scripts and will focus particularly on the nature of the deviations. Research will search for emerging patterns in the interviewer language and behaviour. An attempt will also be made to discover if the findings are marked by either interviewer gender or the school type. The findings will be discussed in light of the theory proposed in chapter one, along with implications for testing practices, interviewer training and test development.

Conclusion will revisit the research hypotheses, summarise the implications of the current dissertation and propose directions for further research.

1. THEORETICAL BACKGROUNG TO MEASURING LANGUAGE PROFICIENCY AND INTERVIEWER BEHAVIOUR DURING ORAL PROFICIENCY INTERVIEWS

1. 1. MEASURING LANGUAGE PROFICIENCY

1. 1. 1. Models of Language Proficiency

Language testing presupposes that the test developer has a theory about what constitutes language proficiency, also referred to as communicative competence or communicative language ability. These theories are frequently expressed in the form of models and as McNamara (1996:48) points out comprise three dimensions: what it means to know a language (a model of knowledge), underlying factors relating to the ability to use language (a model of performance) and how specific instances of language use are understood (actual language use).

The models and frameworks that thus far have most informed the theory and practice of language testing are those of Canale and Swain (1980), Bachman (1990), modified by Bachman and Palmer (1996), Celce-Murcia, Dörnyei and Thurrell's (1995), and Common European Framework of Reference for Languages (2001).

Canale and Swain's (1980) model of communicative competence comprises two elements: communicative competence per se, including grammatical competence (the knowledge of grammar, lexis, morphology, syntax, semantics, phonology, orthography), sociolinguistic knowledge (sociocultural rules of language use and rules of discourse) and strategic competence (strategies to compensate for breakdown in communication); and actual communication. A further element – discourse competence – was added to the first section of the model later (cf. Canale 1993). Canale and Swain make a distinction between communicative competence and communicative performance i.e. actual communication (cf. 1980), but the latter section of their model has not been further elaborated. What the emergence of the model meant for testing purposes was that 'tests should contain tasks that require actual performance as well as tasks or item types that measure knowledge… This is a theoretical rationale for the view that pencil and paper tests alone cannot directly indicate whether a language learner can actually speak or write in a communicative situation' (Fulcher and Davidson 2007:39). Fulcher and Davidson further credit the Canale and Swain's model for allowing discreet point testing as part of communicative competence testing, and for permitting 'a development of criteria for evaluating language performance at different levels of proficiency' (ibid). The above model is praised for its simplicity and ease of application by numerous researchers (cf. Bagarić and Mihaljević Djigunović) who quote and refer to it in spite of the emergence of more recent and more sophisticated models.

Bachman's model of communicative language ability (1990:87) incorporated the second and foreign language acquisition research findings of the intervening 10 years since the publication of the Canale and Swain's model. Bachman defines language competence as being composed of organisational competence and pragmatic competence. Organisational competence has two components: grammatical competence (including vocabulary, morphology, cohesion, syntax and phonology/graphology) and textual competence (including cohesion and rhetorical organisation). Pragmatic competence also consists of two elements: illocutionary competence (including ideational functions, manipulative functions, heuristic functions and imaginative functions) and sociolinguistic competence (including sensitivity to dialect or variety, register, naturalness and cultural references and figures of speech). Communicative language ability is described in terms of five components: knowledge structure, language competence, strategic competence, psychophysiological mechanisms and the context of situation (1990:85). The components of language competence have already been listed above. But it is the strategic competence, made up of assessment, planning and execution, that is the actual generator of speech (1990:106). This model was later modified to include 'topical knowledge' instead of 'knowledge structures', defining strategic competence as a set of metacognitive strategies and introducing affective factors in the model (Fulcher and Davidson 2007:45). Compared to the earlier model, the current model offers a much more detailed description of the language competence as well as a mechanism of how language knowledge is implemented in communication. The implications of Bachman's model to test development concern primarily differentiation in task design and are best expressed in the ALTE guidelines for item writers: 'one can distinguish between two kinds of tasks. Because they demand an immediate response, oral tasks make it necessary for the learner to acquire some forms of routinised communication, and to develop ways of dealing with breakdowns in communication. Written tasks, because they do not demand an immediate response, may include, or even demand, conscious planning. Self-assessment and the learner's own development of control over the learning process are closely tied into the growth of strategic competence. As far as testing is concerned, it may be possible to structure tasks set in a test of speaking in such a way as to make it likely that the candidate will have to resort to compensation strategies' (http://www.alte.org/downloads/index.php).

Celce-Murcia, Dörnyei and Thurrell's (1995) model differs from that of Bachman's in that if the latter was primarily developed with the language testing purpose in mind then Celce-Murcia, Dörnyei and Thurrell set out to generate a model that would be instrumental for syllabus design (Fulcher and Davidson 2007:47). Their model consists of five components: linguistic competence, actional competence, socio-cultural competence, discourse competence and strategic competence. The distinguishing element of their model from the previous ones is the inclusion of what has been labelled as 'actional competence' and is defined as knowledge to interpret 'communicative intent by performing and interpreting speech acts and speech act sets' (Celce-Murcia et al 1995:9). The authors give a detailed description

of the essence of each of the components maintaining that it is the discourse competence that is at the heart of the communicative competence. This in its turn is fed by linguistic competence, actional competence and sociocultural competence, with strategic competence affecting each stage of the interplay. Being guided by the above model in test development has implication to test design and validation in terms of content and means of testing.

Common European Framework of Reference for Languages: Learning, Teaching, Assessment (2001) is probably the most influential test development document in Europe at the moment. Although seminal for test development, the document is labelled a framework rather than a model. The difference between the two can be understood in light of Chalhoub – Deville (1997) distinction, who defines models as abstract and theoretical theories of second language communication ability and frameworks as definitions of particular skills that have been chosen from the model to be tested. Although the main focus of the above document is on particular skills, it does however, offer a brief theory of communicative language competence that the framework draws from (CEFR:13–16). Communicative language competence is seen as comprising linguistic, sociolinguistic and pragmatic components, each incorporating knowledge, skills and know-how. The competence is 'activated in the performance of various language activities, involving reception, production, interaction or mediation'(2001:14). The activities in their turn are set in particular contexts, which have been divided into four: public, personal, occupational and educational. The model does not explicitly mention strategic competence. It is referred to, however, in the section where the performance of tasks is discussed requiring 'use of strategies' (2001:15). Compared to the models above, the one offered in CEFR is theoretically perhaps the least elaborate. The value of the document lies in its function as a blueprint for making decisions about types and kinds of assessment and evaluation, designing marking systems and rating scales, evaluating tasks, etc.

Having a theory about what needs to be tested if we test communicative language ability is essential for developing an informed language assessment tool. Language testing research views models as 'theoretical anchors for describing which features [of language competence] are relevant for the practical purpose for which the test is being used' (Luoma 2004:107).

1. 1. 2. The Concept of Validity

Having established the theoretical basis, a key concept while constructing and evaluating an instrument for measuring language competence is its validity. Validity, the most important test characteristic, is usually defined as 'the extent to which the inferences or decisions we make on the basis of test scores are meaningful, appropriate, and useful' (American Psychological Association 1985 qtd. in Bachman 1990:25). For the language testing development purposes, the test should be constructed so that it measures the language competence with the results not being

affected by errors of measurements or other factors accompanying the measurement process. Traditional validity theory, first outlined by Chronbach and Meehl (1955) divided the validity into criterion-related validity (comprising predictive validity and concurrent validity), content validity and construct validity. Alderson et al (1995) divide the concept of validity into internal and external validity. Internal validity would encompass face validity, content validity and response validity, and external validity includes concurrent validity, predictive validity and construct validity. Research of that time saw the validation of a test as consisting in producing evidence to show that the respective types of validity could be accounted for. Other test characteristics – reliability, practicality and test impact – were viewed independently, but seen ultimately to contribute to particular aspects of test validity (cf. Bachman 1990, Weir 1990, Alderson et al 1995).

The concept of validity has since been reviewed and, relying on the views expressed in the work of Messick (1989), is now viewed as 'an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on the test scores or other models of assessment' (1989:13). Messick presented a progressive matrix of validity facets, consisting of two facets: 'the source of justification' (1989:20) – either evidence or consequence; and the 'function of outcome' (ibid) – test interpretation or test use. Validation of a test according to this view would mean providing evidence to support the inferences that have been made based on the test score and the uses that are made of the score. Reliability, practicality and impact judgements are subsumed into the process as parts of the continuous validation process. Bachman (1990) divides the evidence into three types: content relevance, criterion relatedness and meaningfulness of construct but agrees with Messick that rather than viewing them as discreet components of validity 'they can be more appropriately viewed as complementary types of evidence that must be gathered in the process of validation' (1990:243). For the purposes of current research, validity concerns will be central while discussing the usefulness of the national examination in the English language during its evolution since its introduction.

1. 2. INTERVIEWER VARIABILITY AS A VALIDITY CONCERN

The second main angle the current research takes has been derived from the emerging research, conducted by McNamara (1997, 2001, 2003), A. Brown (2003, 2005) and A. Lazarton (1996, 2002) among others, that has stopped viewing language testing as allowing language proficiency to be displayed during test performance. Instead, language testing is seen as a social practice that 'constructs the notion of language proficiency' (McNamara 2004:339). Such an approach has its origins in the notion of performativity developed in the work of Judith Butler (1990; 1993) and in the dissatisfaction with the earlier definition of relationship between the language performance and the language competence. Bachman (1990) shows that language competence

cannot be directly derived from the language performance during a language test but may for example be affected by the test developer's understanding of the construct (1990:32) and the test method (1990:225). McNamara (2001) emphasises the need to 'consider the complex social construction of test performance, most obviously in the case of interactive tests such as direct tests of speaking' (2001:337) and suggests that the candidate proficiency is co-constructed by a number of participants, for example the communicative partner, rater, test designer (2001:338). He alerts research to resort to 'discourse analytic techniques to reveal the jointly constructed nature of performance in face-to-face oral tests' (2001:340). This prompts us to take a closer view of the role the interviewer as a participant plays in testing speaking.

1. 2. 1. Oral Proficiency Interview as an Assessment Tool

An Oral proficiency interview (OPI) is a widespread technique of evaluating candidate's speaking ability where the candidate engages in a conversation-like activity with the interviewer. It is a 'structured encounter conducted for measurement purposes' (Fulcher 2003:79), where contributions to the interaction are made by both parties – the candidate and the interviewer – but where it is the interviewer who leads the interview Although the interview mainly consists of question and answer sequences, other types of tasks, e.g. picture descriptions, role-plays, etc., can be built into the interview.

While discussing the interview as an oral proficiency assessment tool, Cohen (1994) relies on an extensive study carried out by Perret (1990) who commends the oral interview for its high face validity as an elicitation device (262 in Cohen). Perret criticises the OPI on several accounts, though. On the one hand, he finds that 'the range of linguistic phenomena that the interview can elicit is limited' (ibid), allowing the assessment of 'phonological …lexico-grammatical, and certain discourse aspects, [but not] control over topic or text type, interactive aspects of discourse such as speech functions or exchange structure, or use of language in other situations.' (ibid 263). Kormos (1999) reaches a similar conclusion saying that 'one problem with the most commonly used forms of oral language tests – the oral proficiency interviews – is that they are unequal social encounters, thus they inherently resemble interviews rather than natural conversation. The traditional interview format of language proficiency exams might prove to be an adequate means of measuring linguistic (grammatical, lexical, etc) competencies; nevertheless, several researchers (e.g. Bachman and Savingion 1986, Bachman 1988, van Lier 1989, Lazarton 1992, Young and Milanovic 1992, Young 1995) argue that it does not create a situation in which conversations can occur' (164).

Awareness of the impact that the interviewer has on the progression of the oral proficiency interview emerges in the testing research literature in the 1990s. Young and Milanovic (1992) study the interviewer and candidate roles and find that 'the two parties made very different contributions to the discourse, with the examiner exerting

a controlling influence and the examinee having a more reactive role' (1992:403). They study the OPI from the point of view of interactional contingency, goal orientation and dominance. They find a higher level of examiner control over the discourse during the OPI, which manifests itself in the fact that the topics initiated by the examiner persist longer than the ones initiated by the candidate (1992:416). Although the examiner readily gives the floor to the candidate and the candidate seems to speak twice as much as the interviewer thus dominating the interview, the interviewer still controls what is being talked about and for how long, thus demonstrating a higher level of goal-orientation and dominance (ibid). Perret focuses on the interviewer's role, which is 'to encourage the interviewee to speak' (Cohen 263) and alerts the reader to the fact that the interviewer is usually a virtual stranger to the speaker and, what is more, has the role of considerable power, affecting the candidate's rating (ibid). Louma (2004) corroborates the findings by saying that 'the interlocutor initiates all the phases of the interaction, and asks the questions, whereas the role of the examinee is to comply and answer' (2004:35). Studies note the asymmetrical character of the OPI, where the interviewer is far more in control of the interview than the candidate. Kormos (1999) in her research discusses the participants' rights during a conversation as opposed to the OPI and finds that during the interview, 'it is the interviewer's right to open and close the dialogue and to ask questions by which he or she introduces new topics, whereas in a conversation these rights are shared by both participants. Participants in conversations are also entitled to reject or ignore a new topic, that is, not to ratify it' (1999:170), whereas during the OPI, they are not. She thus comes to the conclusion that 'in an interview situation, one party, the interviewee, is deprived of his or her rights but is heavily burdened by the duties' (ibid). Fulcher and Reiter (2003) endorse this by stating that 'in most situations where speaking is tested there is social distance between the test taker and the interlocutor, even when the test taker is asked to engage in role-play where the characters are meant to be social equals' (2003:330). According to their estimation, task difficulty is affected by social distance between the interlocutor and the candidate, the amount of authority or social power between them, the degree of imposition of the utterance and a cultural factor (2003:331). O'Sullivan (2002) has found research to demonstrate that the candidate performance may be affected by the interlocutor's age, interaction style, language level, personality, sex and status.

1. 2. 2. Interviewer Behaviour

The discussion of interviewer behaviour during oral interviews in the relevant research literature takes the form of discussing the phenomenon known as interviewer accommodation. In their 1991 article, Ross and Berwick call it a phenomenon little investigated but point out that interviewers have a repertoire of strategies to accommodate low proficiency candidates (quoted in Cohen 268). Their subsequent work (1996) outlines a whole array of accommodation features: display questions, a comprehension

check, a request for clarification, or-questions, fronting, grammatical accommodation, slowdown, over-articulation, other explanation and lexical simplification (1996:50–1). In their opinion, 'over- accommodation diminishes the power of the probe, which is intended to push the interviewee's performance of oral proficiency to its limit' (quoted in Cohen 268), and should not be resorted to during the OPI. Malvern and Richards (2002) define accommodation as a process 'by which the speech of participants in linguistic interaction converges or diverges in a systematic way, i.e. how the speech of one person becomes similar to, or different from, that of a conversational partner' (2002:86). They describe convergent accommodation as stemming from the aspiration to social approval or the need for effective communication between speakers. They claim that phenomena like 'foreigner talk' and 'language teacher talk' are instances of accommodation (ibid). Lazarton (1996) maintains that accommodation is a feature of OPIs and in her subsequent work (2002), offers her taxonomy of accommodating behaviours displayed by the interviewers during OPIs: supplying vocabulary, completing responses, rephrasing questions, evaluating responses, repeating or correcting responses, stating questions, drawing conclusions, slowing rate, fronting (or top priming) (2002:128).

Research has not established clearly yet what it is in the candidate behaviour that invites accommodation from the interviewer. There is some evidence (Ross 1992, Ross and Berwick 1992) that accommodative behaviour is related to the candidate's proficiency level.

The above discussion seems to focus mainly on convergent accommodation, i.e. adapting one's speech pattern/ level to that of the candidate, thus in a way stepping down to the candidate level, which seems to be viewed in the negative light, putting candidates in unequal interview situations. There is research to suggest that accommodation, in this case divergent accommodation, would be something that interviewers should engage in during the interviews. Malvern and Richards (2002) investigate accommodation with non-native teachers/examiners and find that 'there appears to be a tendency […] in the context of a public examination conducted by non-native speakers, for each teacher to provide an approximately standard level of language across all the students he or she is testing' (2002:101). While they agree that this may reflect the need to be fair and the examination to be reliable, they suspect that such behaviour may 'fail to stretch the students' (ibid). Thus, by increasing the level of sophistication of the probe, the interviewer may prompt the candidate to demonstrate a fuller spectrum of language ability.

1. 2. 3. The Impact of Gender

O'Loughlin (2002) in his study of the impact of gender in oral proficiency testing proceeds from the findings of a large body of research in the field of language and gender to maintain that 'male and female conversational styles are quite distinct … the female conversational style [being] collaborative, co-operative, symmetrical and

supportive, whereas its male equivalent is portrayed as controlling, unco-operative, asymmetrical and unsupportive' (2002:170). Proceeding from this assertion, he finds it necessary to investigate if gender differences could affect test fairness, i.e. 'whether clearly distinct styles are consistently evident for male and female interviewers' (ibid). His research is made all the more pertinent by the fact that prior studies seemed to have indicated a gender effect of some sort on the interview results, but the results appeared quite contradictory. O'Loughlin had come across studies where higher scores had been achieved with female interlocutors and other studies where the scores were higher with male interlocutors and yet another study where the score was higher 'when paired with an interviewer of the same gender' (2002:171). His study, which involved 16 students and 8 interviewers (4 female and 4 male), indicated that 'there was limited use of overlaps, negligible use of interruptions and widespread use of minimal responses in the interviews, [but] the use of these features did not appear to follow any clear gendered pattern (2002:189). He comes to the conclusion that there may be other interviewer characteristics – the professional orientation of the interviewer, interviewer training – that affect the interviewer behaviour more than gender (2002:190). Lumley and O'Sullivan (2005) study the gender effect during a tape-mediated test and find 'only limited evidence that the gender of the hypothetical interlocutor in a tape-mediated test plays much of a role, although this is apparently not always the case, and it cannot reliably be predicted' (2005:432). Their research seems to indicate that there was a link between the topic and the gender of the interlocutor, i.e. 'when required to talk about a topic they were unfamiliar with to a (hypothetical) foreign male, this [was] more face-threatening than showing their ignorance to an absent male' (ibid). They call on a more thorough research of the interplay of task, topic and interviewer/ candidate gender during an OPI. Brown and McNamara (2004), having reviewed a number of studies on the role of gender in speaking assessment, emphasise the complex nature of its effect and claim that the studies so far 'do not support any simple, deterministic idea that gender categories will have a direct and predictable impact on test processes and test outcomes' (qtd in Davies 2009, 369).

1. 2. 4. Interviewer's Proficiency Level and Personality

Interviewer's proficiency can be defined in terms of his/ her language proficiency but also as his/ her proficiency as an interviewer. Brown (2003) quotes earlier studies where the interviewer competence has affected the raters' evaluation of the interviews. She quotes Morton et al (1997) and McNamara and Lumley (1997) studies who both found that the raters gave a higher rating to the candidate when they perceived the interviewer to be less than competent during the interview (1997:3). McNamara (1996) reported a 'significant and a consistent effect for competence' (1996:243) noting that 'the effect was general and not restricted to a single rater. The effect was rather large, about 0.7 of a score point in raw score terms; put another way, in the most extreme

case a candidate would have a 50 per cent chance of reaching a required score where the interlocutor was perceived as being very competent would see this chance increased to about 65 per cent if rated by a rater who perceived the interlocutor as being less than perfectly competent' (ibid).

Luoma (2004) asserts that 'the interviewer's proficiency level is often not an issue, but personality and communication style certainly are' (2004:38). A case in point here is Brown (2003) study which finds that a candidate can be assigned different proficiency levels after having been interviewed by interviewers with different communication styles. The candidate obtained a higher ranking having been interviewed by the interviewer who structured topical sequences by establishing topics with closed questions and developed them with open questions. She recycled topics and integrated the candidate's responses into the next question. She reformulated failed prompts and signalled closure of topics explicitly. This interviewer gave frequent positive feedback and made her interest in what the candidate said explicit. The interviewer whose interviewing style yielded a lower score for the candidate had no specific topic establishment or development strategy. The interviewer shifted the topics frequently, showed little explicit interest in the candidate's responses and seldom integrated them in the next questions. The interviewer did not reformulate failed prompts and used closed questions to elicit extended responses, which were mostly misinterpreted by the candidate. The interviewer was noted to give infrequent positive feedback (2003:17). In her later (2005) study, Brown finds that interviewers differ in how they 'deploy topics, their elicitation techniques ….the amount of support they produce, the degree to which they scaffold the task and the 'pitch' of the questions.' (2005:260). Morton et al (1997) note that interviewers differ in their rapport-establishing behaviours: tokens of encouragement, indices of politeness, and back-channelling devices. They find that the differences in rapport are noted by the raters and reflected in the ratings they assign to the candidates. McNamara (1996) maintains that the effect of the perceived level of rapport on the candidate rating is even stronger than that of the interviewer proficiency – 'candidates were favoured by raters if they were interacting with an interlocutor who failed to achieve a good raport, by almost a full score point in raw score terms (alternatively, in the most extreme case, a candidate who would have a 50 per cent chance of reaching a required score where the interlocutor was perceived as having established good rapport, would have this chance increased to about 73 per cent if rated by a rater who perceived the interlocutor as having failed to achieve this.)' (1996:243–4). Brown (2005) argues that 'differences in interviewer behaviour that might on the surface be taken as evidence of the natural variation that occurs amongst native speakers (and therefore evidence of test validity) may […] turn out to be relevant to the construct.' (2003:20), which means that interviewer variability has to be a consideration in the test validation process.

1. 2. 5. Interviewer Training

There are considerably more research results available on the rater behaviour and the effect of training on the rater behaviour. Compared to that, interlocutor behaviour is significantly under- researched. Upshur and Turner (1999) quote Lunz et al (1990) who find that 'training cannot make judges equally severe, but it can increase the consistency with which individual judges rate all subjects' (1990:88). Similarly, they rely on Lumley and McNamara (1995) when stating that 'results of training may not last long after a training session [and] a need for renewed training before each test administration' (ibid).

Fulcher (2003) notes that 'interlocutor training has not been undertaken as seriously, or for as long, as rater training' (2003:149) mainly because the awareness that score variation may be associated with the interviewer has a shorter history than the awareness of the connection between the score variation and the rater (ibid). Fulcher, along with many other researchers (Luoma, Brown, Lazarton, O'Laughlin, McNamara etc.), acknowledges the necessity of the interlocutor training but points out the scarcity of information regarding 'what form this might take' (ibid). Brown (2005) maintains that although there may be a conviction among language teachers that simply by having ample teaching experience, teachers have acquired sufficient interviewing skills, her research demonstrates that 'this assumption is naïve at best; interviewers do appear to interpret their task differently and do differ in their interactional style' (2005:260).

Most of the above studies underline a need for a greater variety of interviewer-related studies to substantiate the preliminary findings of the studies discussed above. Bachman (2002) recommends that research be conducted by 'stating research hypotheses in a way that explicitly specifies the effects of the various factors that can affect test performance' (2002:469) and that includes descriptions of features of interlocutors that need to be considered while designing language competence measurement instruments. Accounting for the need to carry out further interviewer – related research, Luoma (2004) states that 'the upshot for the examination boards is an encouragement to analyse interlocutor behaviour and give […] feedback to ensure fair testing conditions for all examinees' (2004:38).

1. 3 CONCLUSION

The key concepts discussed in the current chapter can be divided into three groups, the concepts related to language proficiency models, those related to validity and concepts related to interviewer behaviour. Insight into these concepts will allow us a more informed discussion of the national examination construct. By discussing the above concepts, we are first addressing the question of what is being assessed when language proficiency is measured. This is done by looking at the models of language proficiency. The dissertation discusses 3 language proficiency models and one framework that seem to be most frequently discussed in the literature discussing

the definition and interpretation of the concept of language proficiency. The English language national examination is first and foremost connected with the language proficiency philosophy expressed in the Common European Framework of Reference for Languages, where the language competence is seen to be comprised of general competence and communicative competence. The speaking proficiency in the CEFR is discussed in terms of spoken production and spoken interaction. Consequently, the national examination of the speaking test is designed to comprise tasks to elicit both oral production and interaction.

The second set of concepts, those relating to validity, approach the research task from the perspective of how the assessment should be done – so that the instrument that we apply, the procedure we follow and conditions we provide would yield us results that would be meaningful from the point of view of a candidate's language proficiency. The discussion of validity allows us to specify what we mean when we discuss national examination validity and what sort of evidence can be produced (the inferences that have been made based on the test score and the uses that are made of the score) in order to validate the test. With the national examination in the English language, validity judgements are made based on the construct that is reflected in the respective skills tests, rating and monitoring procedures, test construction practices and presentation and application of test results.

The third set of concepts, those related to the interviewer behaviour, address the central variable of investigation in this dissertation. By looking at previous research into the various aspects of interviewer behaviour, it can be specified how interviewer behaviour can challenge speaking test validity and as the speaking test is a part of the proficiency examination – the overall validity of the national examination results.

Having a theory about what needs to be tested if we test communicative language ability is essential for developing an informed language assessment tool.

2. HISTORICAL BACKGROUND

The tradition of systematic nationwide standard language testing in Estonia is not long and can be traced back to the regaining of independence and the developments that started in the Estonian education after that. The first official nationwide tests at the end of the upper-secondary school were launched in 1997, when national examinations both at the end of basic school (form 9) and upper-secondary school took place. Upper-secondary/gymnasium exams were available in the students' mother tongue (March 25), English and German (May 3), history (May 10) and chemistry (May 31). Pilot exams were also available in mathematics (May 17) and biology (May 24) (NE 1997:4). Since then, national examinations have been an inseparable part of the general education system in Estonia.

2. 1. THE PREPARATION PERIOD

The launch of the national examination was preceded by a fairly long preparation period. The Ministry of Education 1997 report 'Riigieksamid 1997' sets the start of the period on Jan 31, 1993 when the then Ministry of Education and Culture regulation 6 'Põhikooli ja gümnaasiumi õpilaste järgmisesse klassi üleviimise, lõpueksamite korraldamise ja kooli lõpetamise kord' divided exams into national and school examinations and appointed the National School Board to define the essence of national examinations, the form they should take and the time of their implementation (NE 1997:4). It took almost a year of planning and discussions before on December 23, 1993 a decree was issued by the chancellor of the ministry to set up a working-group to put together a relevant project. The effort to launch a national test for upper-secondary/high-school/gymnasium graduates was not restricted to the English language only, but involved all languages taught in Estonia (Estonian, Russian, English, German, French) and also sciences. A lot of general training for test developers at the start of the project was conducted to all subject specialists together, but to date, all subjects-specific national examination development groups are working fairly independently. (Alas, Liiv 2009:20) It is interesting to note that although the foreign language teaching spectrum was quite wide in Estonia at that time (alongside with English, the Russian, German and French language teaching had well-established traditions), it was the British model that was chosen as a model for the Estonian national examination system. The English language national examination working-group (Viive Lätt, Kristi Mere, Eve Sass, Kaarin Truus and Ülle Türk) was instrumental in assembling the team to put together the first pilot tests in the English language and subsequent first official national exams in the English language. The principle of assembling the team was to incorporate language teachers as well as ministry and university representatives in the examination development process as the national examination was to serve many purposes and it was important to have the most important stake-holders involved.

The working-group submitted a draft for the national exams to the ministry in June 1994. It was also in 1994 that the Ministry of Education and Culture started cooperating with the British Council and the University of Lancaster in order to prepare for the introduction of national exams in the English language. The above-mentioned working-group of language teachers, university and ministry representatives went on to participate in an in-service teacher education course at Lancaster University, a course that focused on general conventions of foreign language testing of the time in Western Europe and was especially designed to get the national examination project in Estonia off the ground.

2. 2. THE BASELINE STUDY

As a precursor to pilot tests and as part of the Lancaster University training course, the team conducted a baseline study (Lätt et al 1994) in Estonia, investigating the teaching and testing situation in Estonia at the time (in 1994), studying the Tartu University entrance exam requirements in English as well as the school-leaving examinations (11 different school-leaving tests were investigated).

The analysis of the teaching situation of the time emphasises the increase in the importance of teaching the English language stating that 'English supplanted Russian as the most important and widely-taught foreign language [in 1993]'(Lätt et al. 1994:3), but maintains that 'there is considerable variation in both the number of English classes the students have to attend and in the qualification of the English teachers' (ibid 8). With regard to the number of English classes a week, the research established 8 different frequency patterns with the number of classes varying from as few as 2 to as many as 8 per week (ibid 4). Little data exist about the teacher qualification situation apart from the Ministry of Culture and Education Booklet of Educational Statistics No 4, 1994: Foreign Language Learning. In this booklet, V. Rajangu reports the qualifications of full-time English teachers of the time. With regard to the then 998 full time English teachers in the Estonian schools, 73.1 per cent are ELT specialists, 14. 5 per cent have a higher education diploma in another field and 12.3 per cent are trained in either vocational or secondary education (Rajangu, 15). This, however, is not a complete picture of the English teachers in the classrooms. Because of the high demand for English classes, a substantial amount of the teaching load is carried by part-time teachers and 'more often than not, they are not qualified to do so' (Lätt et al 5). The study points out several positive developments that manifest themselves in the English language teaching scene in 1993: it stresses the teachers' enthusiasm about their 'freedom to decide what to teach and what materials to use' (ibid 2) and notes 'a shift towards communicative language teaching (ibid 2) with the help of more modern textbooks obtained from the west, relying on an increased teaching competence gained from widely available in-service courses either in Britain or Estonia or both (ibid 3). By the same token, the study observed serious problems with educational legislation (a lack of a legally adopted curriculum, controversies in

the annually published Suggested Ministry Guidelines with regard to school-leaving examinations and inadequate textbook provision on the part of the Ministry of Culture and Education), which lead to 'the teachers despair over the so-called freedom and their desire for a more unified system' (ibid 7).

The baseline study also evaluates the English language testing situation in Estonia at the time. School-leaving examinations in the English language were optional for the students and composed by individual schools on the basis of Suggested Ministry Guidelines. The Guidelines for 1994 recommended a 3-part English examination, with task 1 focusing on reading and summarising either in English or Estonian a 150–200-word unknown text; task 2 requiring an oral discussion of one of 23 given topics; and task 3 requiring completing a grammar exercise (either a cloze test or a discreet item test, the structures being listed in the guidelines) (ibid 9). Examining the school-leaving examinations that had been administered by different Estonian urban and rural upper-secondary schools during the two years preceding the study, the study finds that although Suggested Ministry Guidelines exist for conducting school-leaving examinations, 'none of the teachers observed completely followed the [Guidelines]' (ibid 16). Rather, the final examinations attempted to replicate the entrance examinations the students would have to take were they to attempt entry into either Tartu University or Tallinn Pedagogical University (currently Tallinn University) English departments. A huge washback effect was created in 1991 in the admission practices in Tallinn Pedagogical University (currently Tallinn University) and in 1992 by Tartu University which decreed that 'a foreign language examination will be the first of the three entrance examinations to all the departments of the university … Those who fail the examination will be excluded from further competition' (ibid 11). Taking that decision, not only did those universities deem the foreign language competence to be of critical importance from the point of view of students' career development, and consequently of great demand by students in both state and private schools, but it was also the testing techniques implemented during the English language entrance examinations at the universities that found their way more frequently to final achievement tests at schools. As the newly introduced Tartu University entrance examination in the English language was a written examination, consisting exclusively of multiple choice items (ibid 12), that particular testing technique was found to be the most 'widely used testing technique' (ibid 14) in the school-leaving examinations analysed. All school-leaving examinations consisted of two parts – an oral examination and a written exam (the latter incorporating besides multiple choice items also gap-filling tasks, translation from Estonian into English, short answer questions, cloze tests and writing a summary) (ibid 14). Although school-leaving examinations were conducted in the English language, their design, choice of content, administration and analysis followed no centralised pattern. Hence the reluctance on the part of the universities to trust results of those examinations and their practice to require participation in the entrance examinations from all candidates. With the school – leaving examinations, the researchers note a lack of proper test specifications, pre-testing, training of item writers, test administrators and test markers as well as monitoring of markers.

The tests are constructed by the teachers who may discuss the items with the colleague when writing the test, but there is no formal procedure for either item writing or test construction per se. No relevant statistics are calculated and the only information reported to the School Board is the number of students receiving a particular grade (Lätt, et al. 17–18). Consequently, no conclusions can be drawn about the validity or the reliability of the tests other than by comparing the results of those students who take the English language university entrance test.

On the basis of the findings and relying on the new-found foreign language testing theory acquired during the in-service training course at the Lancaster University, a detailed set of test specifications was produced (cf. Lätt et al 20–32) for an envisaged first version of a national English examination. 'The Specifications for the Year 12 Examination' was a document developed in conjunction with the baseline study and outlined the proposed test in great detail. The specifications consist of 5 chapters, delineating first the general statement of purpose, the target language situation, a theoretical framework, assessment objectives and finally the form of examination (ibid 20–32). This is the first attempt to base the English language testing practices on a solid theoretical foundation (cf. ibid 21), to determine the test-type that is being constructed (a proficiency exam) (ibid 20), the expected level of the candidate attempting the test (Threshold 1990) and to outline in great detail the number of papers it includes (5 – listening, reading, speaking, writing, language use) (ibid 22). Each paper is subsequently described in terms of abilities tested within that paper, the nature of texts used, text types, task types and marking principles. It is interesting to note that the working-group's initial plan was to offer year 12 examinations (national examinations) on 2 levels – ordinary and higher level (ibid). No other reason apart from allowing 'candidates of all levels of competencies to demonstrate their abilities in assessment objectives' (ibid) was offered for that decision and looking ahead, it becomes evident that this plan proved to be somewhat immature. Neither the pilot tests nor the first official national exams in the English language implemented that suggestion. The decision may have been prompted by the widely differing amounts of input students in Estonia had in terms of the number of lessons available to them in different schools. Thus it may have seemed feasible to offer a simpler version of the test to those with only 2 hours of English a week and a more advanced level for those with more.

The national exam in the English language was designed to have multiple functions from the start. On the one hand, it was planned as an instrument to evaluate students' foreign language proficiency level at the end of upper-secondary school, to allow comparisons between students. While doing that, educators were able to see to what extent the students had acquired the material included in the upper secondary school/ gymnasium curriculum. At the same time, relying on the examination results, inferences could be made about the effectiveness of foreign language instruction. On the other hand, the Ministry of Education and Culture was interested in obtaining a comprehensive overview of the standard of education provided in different schools in Estonia, and administering a national examination provided a tool for them to do so. In addition to the aforesaid, national exams were also recommended from the start

as being simultaneously entrance exams to the universities. Up until 1996, the departments of universities in Estonia set entrance exams to those who applied to that particular specialty. Thus the students finishing upper-secondary school/gymnasium and entering a university were exposed to between 4 to 10 examinations depending on the specialty the student chose. The national examination system aimed to reduce the number and set uniform conditions to all school leavers and university applicants. It was on November 1, 1996, that the state universities adopted a resolution about accepting national examination results as proof of student proficiency and utilising them in their enrolment procedure. (NE 1997:5).

The first national examination took place in the students' mother tongue (Estonian or Russian) on May 29, 1995. Other subject groups were working on their own pilot exams. That same year, the first pilot exam took place in English (the second pilot followed the next year). Participation in it was voluntary. The next year, June 12, 1996, the Ministry of Education, the local education boards and the School Headmasters' Union adopted a resolution with regard to conducting official national exams in 1997 in five subjects – mother tongue (Estonian or Russian), English, German, history and chemistry – and to conduct pilot exams in biology and mathematics (NE 1997:4).

2. 3. PILOT EXAMINATIONS 1995 AND 1996

In order to develop a full-fledged test from the specifications worked out by the working group, the test-developing team was substantially increased to include about 60 item-writers, 18 exam developers and 69 test markers (Personal data files of Kristi Mere). The national examination in the English language was piloted twice – in the spring of 1995 and 1996. Participation in both examinations was voluntary and the students who were not satisfied with the results obtained during the pilot examination could take the so-called final school examination (NE 1997:4). The exam population varied but showed an increase in interest on the part of the students. In 1995, 222 students took the pilot examination, whereas the number grew to 1304 in the following year. The speaking section was piloted with 285 students in 1996 (NE 1997:22). All 6 republican towns and all 15 counties were represented.

The structure of the pilot examinations is presented in the table 1 below.

Table 1. Pilot examination structure.

Skill Section	No. of tasks	Items	Max points	Time (minutes)
Lg. structures	4	17+5+17+20	59	55
Listening	3	15+15+17	47	25
Reading	4	11+4+7+10	32	40
Writing	2		12+16=28	60
Speaking	2		16	12–15

Origin of data: Personal data files of Kristi Mere

It can be observed that there is variety in the number of tasks the students are expected to complete within each skill section as well as the number of points they will receive in each case. The sections are unevenly weighted, with the maximum number of points given for the language structures section. It is interesting to note that the greater number of points can be earned from the sections that resort to indirect language proficiency measurement (language structures, listening, reading), while the skills that can be measured directly (writing and speaking) yield a considerably smaller number of points. If thus implemented, the probable impact on the classroom practices would be that the receptive skills and language structures would be practiced more than productive skills by teachers and students, as a successful completion of the latter would guarantee a higher number of points at the examination. It should be noted here already that the weighting of different sections is going to be altered when the first official national examination is administered.

Although the aim of pilot testing had been to 'estimate the difficulty level of the examination tasks and check their correspondence to the general level of the school leavers as well as monitor the process of exam administration, marking of papers and results' presentation' (Personal data files of Kristi Mere), the all-Estonian pilot testing allowed the Ministry of Education to draw other conclusions that had not been possible earlier. Comparisons were established between the results of the whole test and different skills sections, the results of 1995 and the 1996 overall results and the skills sections results. Difficulty levels were calculated for all the tasks within skills section alongside with the correlation indices between particular sections.

For example, the 1996 pilot test results can be summarised as follows:

Table 2. Pilot test results 1996.

	Language structures	Listening	Reading	Writing	Whole test	Speaking
Mean	66.3%	78.5%	65.9%	65.4%	69.4%	76.9%
Maximum	100%	100%	100%	100%	94.4%	100%
Minimum	10.2%	27.7%	3%	0%	21.7%	31.3%
Range	89.8%	72.3%	96.9%	100%	77.7%	68.7%

Source: ibid

The analysis finds a marked difference between the results of the 1996 and the 1995 pilot test, noting a considerably lower average proficiency level in all skills except listening in 1996 (cf. in 1995 the mean scores were calculated as follows: language structures – 69%, listening – 72%, reading – 78%, writing – 65.4%. Speaking was not tested in 1995) (Personal data files of Kristi Mere). The 1995 pilot exam results paint a rather glowing picture of the average English language proficiency level in Estonia. 'In a normal distribution … we know that 50 per cent of the scores are below the mean, or average, and 50 per cent are above' (Bachman 1990:73). The distribution of scores here seems to be negatively skewed, judging by the mean, which indicates

'scores clustering at the mid and high score levels' (Bachman, Kunan 2005:43), not representing the normal distribution. The fact that the distribution is scewed may indicate a problem with either the test or the test population. The test developers themselves ascribe the lower results to the possibly increased difficulty level of the test or the better general proficiency of the test takers in 1995 (Personal data files of Kristi Mere). It could also be, however, that the somewhat more realistic results of the 1996 test were achieved because of a bigger and more representative test takers' population.

Comparisons were also established between schools depending on the number of classes of English offered to students, and between the results of students who had started English as their first foreign language and those who had started it as a second or third foreign language. In addition to that, schools were compared based on the medium of instruction (Estonian or Russian), their geographical location (e.g. Tallinn schools compared to Tartu schools, urban schools compared to rural schools). All these factors seemed to have a bearing on the students' examination results, most notably the number of English classes a week (to make the average result at least 4 classes of English a week seem to be necessary), and the location of schools (urban school students did generally better than rural school students) (ibid).

An additional correlation was established during the pilot testing of 1996 that further corroborated the need for a uniform, centrally administered proficiency evaluation system. Before the pilot exam took place, teachers of the participating students were asked to estimate their students' probable result during the examination. Correlation was then established between the teacher's estimated result and the actual result attained during the examination. The overall correlation was a low 0.49% (ibid) pointing to either the teachers' inability to correctly evaluate their students' proficiency or their desire to boost their students' result (by the same token sending a message about his/ her own good teaching ability). About 60% of the teachers achieved a fairly good correlation varying between 0.6 and 0.9 (ibid). Such huge variation, some teachers estimating that all his/ her students attain grade excellent, others applying different criteria of evaluation depending on how many classes of English their students have in a week, clearly showed that unless an instrument of assessment was introduced that allowed fairer and more uniform evaluation practices, the students proficiency results reported on their graduation transcript would continue to be obscure at best and completely arbitrary at worst.

Pilot test analysis led the test developers to the following conclusions and action plan for the up-coming 1997 national examination:

1. The difficulty level of the pilot test was correctly estimated for the Estonian upper-secondary school/gymnasium leavers.
2. Having a centralised team of markers for the examination was well justified: all marking was completed based on a common marking system added to the reliability and comparability of marking and consequently to the overall test reliability. This is especially relevant with regard to subjectively marked sections of the examination – writing and speaking.

3. The listening section of the test would have to have an increased level of difficulty.
4. A radio station ought to be used for the listening section that would be available all over Estonia (during piloting, several schools had had to resort to reading out the tape scripts by one of the members of the examination committee, as the relevant radio station could not be tuned into).
5. In all sections the tasks would have to be sequenced in the increasing order of difficulty.
6. It is imperative that the examination administration be monitored by external invigilators in every school.
7. Writing section will have to be double marked.
8. Experts will have to be involved in establishing the pass marks.
9. More interviewers and assessors need to be trained for the speaking section of the examination.
10. Time limits for each section will have to be more clearly established and followed. (ibid).

Considering the above, it can be seen that on the one hand the pilot testing further corroborated the need for nation-wide, centrally developed and administered proficiency testing. On the other hand, the practicalities relating to national examination development, administration and analyses posed challenges that needed to be addressed if valid and reliable results were to be attained.

2. 4. THE FIRST NATIONAL EXAMINATION IN THE ENGLISH LANGUAGE – 1997

On October 1, 1996, a joint conference of the Ministry of Education and universities took place as a result of which a long-awaited resolution was adopted to recognise national examinations both as school leaving examinations and as university entrance examinations (NE 1997, 5). In order to fully concentrate on the development of a national examination and qualification evaluation system, on October 9, 1996, the Ministry of Education signed a further regulation number 182 to establish as of January 1, 1997, a National Examination and Qualification Centre (NEQC) (ibid). On November 25, 1996, regulation number 217 confirmed the statute of the newly-established centre:

1) to set general education schools national examinations and vocational schools state graduation exams in accordance with the regulations set by the ministry;
2) to regulate and coordinate vocational qualifications, state standards and development of respective curricula in vocational training;
3) to set level test papers and tests in general subjects based on the ministry and school requirements;

4) to coordinate adult state language education in the whole country;
5) to coordinate administration of state language examinations;
6) to issue diplomas and certificates related to the abovementioned activities and keep respective records (NE 1997:5).

An event that put the development of the 1997 national examination on a more solid basis still was the fact that the Estonian government resolution number 228 6.11.1996 confirmed the adoption of a national curriculum that explicitly delineated subject curricula, including one for the English language. The content validity of the national examination could thus be established both relying on the national curriculum as well as the examination specifications specified within the baseline study.

The first national examination in the English language in 1997 had 5 papers. Paper one, reading, had 3 texts representing fiction, newspaper and popular science as text types, and utilising matching and true/false/no information as task types respectively. Paper two, writing, required students to write a letter to a magazine (80–100 words) and an essay (120–150 words). The letter had an advertisement as a prompt and had detailed specifications as to its content. The essay suggested five aspects as subtopics for the essay and required the inclusion of at least 3 of them. Part 3, listening, included three listening excerpts – a story, an airport announcement, a talk about a British institution and had summary identification, gap-filling and sentence completion as task types. Paper four, language structures, comprised 4 tasks and required word-building, editing, inserting the right word in a sentence or deriving the right verb-form. The oral section started with a warm-up and proceeded to a picture-description followed by a role-play.

The discussion revolving around the 1997 national examinations resulted in numerous articles in the main daily and weekly newspapers. Adamson and Kond (NE 1997:5) report over a hundred articles that were published commenting on the examinations. Looking at the feedback concerning the first round of national examinations of 1997, there seemed to be a consensus regarding the necessity of a national system of evaluation. The respondents appreciate the fact that due to the national exams, it is now possible to compare the exam results and consequently the students' proficiency; that the exam resorts to both the direct and indirect ways of testing students' knowledge, and that sub-skills are tested separately (Türk, 1997). The students value fair marking and increased objectivity of marking resulting from multiple scoring. They also consider important that national examinations count as both school leaving and university admission examinations (http://greta.cs.ioc.ee/~opleht/Arhiiv/97Mar21/artikkel10.html). In spite of the general agreement that national examinations are necessary, several educators draw the public's attention to the flaws either in the exam paper or in the examination administration. Regarding the English examination, Läänemets (1997) comments on the generally too difficult level of the whole examination, suggests that a different ordering of tasks could have been more appropriate and complains about typos and an inadequate marking key of the exam, Penjam (1997) writes about substandard

examination preparation materials. With regard to general national examination administration, Reiman (1997) finds the students' examination results to be quite poor, which is seconded by Adamson's (in Liimal 1997) evaluation to the same effect. T. Märja (in Malmberg 1997) notes the teachers' general mistrust of outside assessment and some sources (e.g. Kapp, 1997) point out future efforts that would have to be put in the exam management resulting from numerous incidents of cheating during the examination in certain schools. Thus from the outset, the problems connected with the examination are to do with the examination validity and the reliability of its results.

2. 5. NATIONAL EXAMINATION CONTENT 1997–2008

Although the national examination has been intended as a proficiency examination from the beginning – an examination that is 'designed to measure people's ability in a language, regardless of any training they may have had in that language' (Hughes, 11), the examination also bears some features of an achievement test, 'tests that are directly related to a language course' (ibid, 13), as it tests to what extent the students have mastered – achieved – the goals set for the upper-secondary school leaver in the English language in the national curriculum. As the national exams do not only 'relate to the past in that they measure what language the students have learned as a result of teaching' (McNamara 2000, 7), but more importantly 'to the future situations of language use without necessarily any reference to the previous process of teaching' (ibid), the national examinations are first and foremost proficiency exams. The twofold nature of the examinations is reflected in the relevant educational regulations that govern the examination design and development. The design and development of the English language national exam in Estonia proceed from the Ministry of Education and Science regulation of January 23, 2001 no. 18 "Õpitulemuste välishindamise põhimõtted, riigieksamitööde, põhikooli eksamitööde ja üleriigiliste tasemetööde koostamise, hindamise ja tulemuste hindamise alused" (Regulation 2001). The regulation specifies the purposes of the national exam as follows:

- to evaluate the attainment of the educational goals outlined in the basic and gymnasium curricula;
- to give schools and teachers an opportunity to compare the results of their students to those achieved by other students in the country;
- to steer the educational process through the content and form of national examinations;
- to link consecutive educational levels and stages;
- through external marking, to give feedback to all stakeholders and to allow planning and execution of changes in the national curriculum, textbooks, in-service training of teachers and allow development in the respective areas.

As can be seen, the purpose of the national exam has in broad terms remained similar to its initial envisaged purpose. Consequently, what the exam developers have to constantly be aware of is the enormous washback effect in terms of teaching and testing practices at school and its impact on the stakeholders. "Stakeholders would include the test designers, teachers, students, score users, governments or any other individual or group that has an interest in how the scores are used and whether they are useful for a given context" (Fulcher and Davidson 2007:14). The impact of the exam can be illustrated by looking at the entrance requirements set by particular universities: e.g. out of 59 specialities admitting students to Tallinn University BA level studies in 2008, 24 specified the foreign language national examination result as being of criterial importance during the admission procedure (Alas, Liiv 2009:20–21).

An important change was made in the structure of the national examination starting from 1998. Instead of the reading section, the new exam started with the writing section (75 minutes), the time allotted for which was increased by 15 minutes compared to the 1997 examination. The writing section increased the number of words required from 80 to 100 words with the first task and 120–150 words with the second task in 1997 to 140–180 with the second task in 1998 (NE 1998:25). Although the number of words required with the first task has not been indicated (cf. NE 1998:23) there is reason to believe that the expected length for the first task was also increased, as suggested in the in 1997 description of the following year's national exam (ibid) and specified with the first task in the 1999 examination – 80–120 words (NE 1999:22).

The 1998 examination also introduced a break between the writing section and the following three sections of the examinations: after the break the students completed first the listening section – 30 minutes, the reading section – 50 minutes and the language structures – 40 minutes (NE 1997:23).

Table 3 below illustrates the structure of the current national exam in the English language as it stands today, specifying the number of tasks in each section, the maximum number of points available for that section and the time allotted for the completion of the section.

Table 3. National examination structure.

	Skill Section	Tasks	Maximum points	Time (minutes)
1	Writing	2	20	80
2	Listening	3	20	35
3	Reading	4	20	50
4	Language Structures	4	20	40
5	Speaking	2	20	13–16

The time given for each section has generally remained the same over the years with two exceptions. In 2001, the time for the listening section was extended from 30 minutes for 35 minutes, and in 2006, the time for the writing section was raised from

75 minutes to 80 minutes. Skill sections 1 to 4 are completed consecutively on the same day, with the speaking test taken on the following day. Compared to other skill papers, the speaking test allows the examiner some freedom as to the time within which the test has to be completed. This is done in order to consider the idiosyncrasies of the examinees, allowing for varying rates of response and speech speed.

Test content is one of the many indicators of a test's validity. Though detailed specifications have been developed and a national curriculum exists that clearly states what the students should have mastered by the time they sit for the national examinations, it is not feasible that all the aspects in the specifications or in the national curriculum will find their way to a particular national examination. It is only so much that can be accomplished within the envisaged time. Consequently, only a selection of the content of both the specifications and the national curriculum will find its way into the actual version of the national examination every year. 'For content validity and for beneficial backwash, the important thing is to choose widely from the whole area of content. …Succeeding versions of the test should also sample widely and unpredictably, although one will always wish to include elements that are particularly important.' (Hughes 2004:63). Looking at the content of the respective skills sections in the national examination in the English language in Estonia, it would be interesting to see, to what extent this maxim to ensure content validity is reflected in the particular national examination papers over the years.

From the historical point of view it should be noted that the test development team has undergone three important changes in its membership. The initial team that started its training in January 1995 continued developing the national examination until 2001. That year the whole team resigned and was replaced by a completely new team of test writers, some of whom had worked as item writers before, but the team as a whole lacked systematic training in test development that had been given to their predecessors. A subsequent change occurred in 2006, when consultants were added to the test development team whose function it was to safeguard test security, but also edit the tasks written by the item writing teams to provide consistency in item difficulty.

Below each national examination paper will be discussed in terms of tasks it was expected to contain and the tasks that it actually contained and some conclusions drawn as to the content validity of the relevant section.

2. 5. 1. Writing

Table 4. Task types in national examination writing paper from 1997 to 2009.

Year	Task Type
1997	1. a letter to a pen-friend 2. an essay: What might attract tourists to Estonia?
1998	1. a letter of advice to a friend relying on pictures 2. an essay: What is important when choosing a job?

1999	1. a letter of apology to a friend 2. a report: an evaluation of an English language course
2000	1. a letter to a friend making plans 2. an essay: Mobile Phones: For and Against
2001	1. a letter to a friend about school customs and traditions 2. a report: changes in favourite pastimes
2002	1. a letter of application 2. an essay: The Most Needed Professions in Estonia in the 21st century. Discuss Three Professions.
2003	1. a letter of enquiry 2. a report: an environment project
2004	1. a letter to a friend, making arrangements 2. an essay: having a job while studying.
2005	1. a thank-you letter to sponsors. 2. an essay: If I were the Minister of Education. (3 problems and solutions.
2006	1. a letter of complaint 2. a report: Interest in the school library
2007	1. a letter of advice 2. a report: students' music preferences
2008	1. a letter of protest 2. an essay: Advantages of Going to the University and Going to Work after Finishing School
2009	1. a letter of enquiry 2. a report: students' eating habits

The **writing** paper has two tasks, the first of which is a letter and the second task is either an essay or a report. The expected length for a letter up until 2006 was specified as between 80 and 120 words. In order to avoid awarding similar points for exam responses of substantially differing lengths (e.g. one student writing 80 words and scoring maximum points and another student writing 120 words and also scoring maximum points) the requirement was changed as of 2007 where all the examinees are expected to write 120 words and are penalised if the response is significantly shorter. The required length for the second writing task (essay/report) was set at 200 words in the same year. Here, too, a range (from 150 to 200) was allowed prior to that, which potentially may have given rise to unfair test scores. Another change in this task involved the genre. The national curriculum of 2002 specifies that the 'written competencies of the a gymnasium graduate include the following: writing messages, formal and informal letters; taking notes of information obtained either through reading or listening, fill in forms and questionnaires, write a CV, write descriptive, argumentative and discursive essays, reports and short articles for the newspaper; knows the main principles of punctuation, can appropriately use paragraphs and format tests.' (https://www.riigiteataja.ee/ert/act.jsp?id=12888846 29.06.2009). Year 12 Handbook, which effectively functions as the specifications document for the national examination in the English language, in 2005 still specified the expected text types in three categories: 'public writing, i.e. form filling, formal letters; social writing, i.e. instructions,

notes and messages, postcards and personal letters; study writing, i.e. stories, essays, reports" (Jõul et al. 2005:14). Looking at the table above it can be observed that there has been an increase in the level of difficulty in the first writing task: if between 1997 and 2001 the task invariably was to write a letter to a friend (informal writing), then starting from 2002, with only one exception (2004), the students are expected not to write an informal but a semiformal or a formal letter of different genre (e.g. enquiry, apology, complaint, protest, etc.). Relying on the national curriculum guidelines and the Common European Framework for Reference: Learning, Teaching and Assessment (CEFR) B2 level writing guidelines (CEFR 2001:61–62), the test developers seem to have moved in the right direction keeping the Estonian national examination in the English language more in line with all-European developments. Setting particular task types, e.g. writing notes and message, filling in forms, does not allow students to demonstrate their proficiency beyond the level of B1 (CEFR 2001:84). By the same token, it is only when writing other than informal letters, that the writer will have a chance of demonstrating his/her awareness of the audience and make a clearer distinction between spoken and written language, i.e. establish oneself as a B2 level language user rather than a B1 language user (CEFR 2001:83). As according to the Estonian national curriculum, by the time students finish upper-secondary school/gymnasium they are expected to have reached the B2 level on the CEFR scale (https://www.riigiteataja.ee/ert/act.jsp?id=12888846 29.06.2009), the national examination will have to be set so that it includes tasks that would allow the students to show B2 level proficiency.

The second writing task has either been an essay or a report all through the years, as can be seen in the table above. There are two types of writing that although specified as task types that have to be mastered by the end of upper-secondary-school/ gymnasium, have never appeared as actual tasks in the national examination: a story (cf. Jõul et al 2005:14) and a short article for the newspaper (https://www.riigiteataja.ee/ert/act.jsp?id=12888846 29.06.2009). The exclusion of the two task types may have been prompted by practical considerations: if the students proceed to go on to the university, they will be less likely to need story or newspaper-writing skills (except those few who will continue to study journalism, and even then they would seldom have to write in English) than essay or report writing skills. Another consideration may have been the teachers' own proficiency level in the two areas and the alleged lack of time within the given number of classes to practice story and article writing. A further challenge may have been the task design and marking constraints with regard to both task types. The writing assessment has tried to confine itself to assessing the skill of writing rather than the creativity of students insofar as the two can be separated. With regard to both the story and the short article, it seems that the students who are more creative will have an advantage over the students that has little to do with the writing skill. It may have been any one or all of the above considerations that have so far led to the exclusion of the above-mentioned task types from the national examination. There are other considerations, however, which should warn the test designers against such practice. Alderson et al (1995) in their 'Language Test Construction and Evaluation'

maintain that 'if the test format remains fixed for a period of time, it may have the effect of narrowing the curriculum: not only will the test be confined to those elements that are thought testable and convenient, but the teaching in preparation for the test is likely to become restricted to the sorts of activities and abilities that are tested.' (1995:228). Thus, including only essays and reports as possible tasks in the national examination, will result in teachers spending as much of the time as they have available on honing students' essay and report writing skills, which in effect means that mart of the curriculum does not get taught altogether. Empirical research would be necessary to see how much time is actually devoted to learning to write stories or short newspaper articles. Alderson et al (1995) suggest a remedy for the above situation: 'to avoid such a narrowing, as well as to improve content validity, some testing bodies deliberately adopt a policy of constant innovation each year. For each test administration some part of the test is changed…' (1995:228). This would mean finding a way to incorporate both the story and the short article as alternative tasks for the essay and the report in the English language national examination.

2. 5. 2. Listening

Table 5. Task types in national examination listening paper from 1997 to 2009.

Year	Task Type
1997	1. Identify true statements 2. Complete time-table 3. Complete text
1998	1. Multiple Choice (MC) 2. Complete message forms 3. Complete text
1999	1. Match title and message 2. MC 3. Complete notes
2000	1. Ordering 2. Complete message forms 3. Complete text
2001	1. Match title and message 2. Complete notes 3. Complete txt
2002	1. Complete table 2. Complete text 3. Complete text
2003	1. MC 2. Complete form 3. Complete text
2004	1. Complete form 2. MC 3. Complete text

2005	1. Complete notes 2. MC 3. Complete text
2006	1. Complete form 2. Complete text 3. Complete text
2007	1. Complete notes 2. Complete text 3. Match title and message
2008	1. Match title and message 2. MC 3. Complete notes
2009	1. MC 2. Complete notes 3. Match title and message

The **listening** comprehension paper has consistently had three tasks that employ text types such as public announcements, instructions/ directions, interviews and conversations between two or more people, mini-lectures, telephone messages, radio programmes, etc. (cf. NE 1997–2009). In this respect, the particular national examination papers follow the specifications for text types (Jõul et al 2005:37). As far as tasks are concerned, every consecutive task is intended to have an increased level of difficulty, which is decided by the pilot stage results. The 2005 Year 12 Handbook (specifications) lists seven task types for evaluating listening skills: questions (yes/ no, multiple choice, short answer questions), matching, ordering, following instructions, note taking, information transfer, completing (NE 2005:38–39). The national curriculum specifies that the listening competencies of a gymnasium graduate include the following: understanding everyday conversational language of different speakers and messages transmitted over the telephone on condition they are delivered in a language variant close to standard language; being able to follow radio and TV news and announcements to get necessary information, can distinguish between different tones and attitudes, notice emphasis and thought units; being able to guess the meaning of unknown words from context and given elements; knowing the meaning of more frequent international words in his/her native language and being able to utilise this knowledge when listening to a foreign language text; being able to distinguish between detail and sequence events; being able to follow a short lecture (5–10 minutes) and glean relevant information from it (https://www.riigiteataja.ee/ert/act.jsp?id=12888846 29.06.2009). The table above shows a fairly good coverage of the task types mentioned in the specifications: with the exception of note taking, all the task-types have featured at least once. Completing the notes, as it features in numerous national examination listening papers is not equivalent to note-taking as defined by the testing theory – 'candidates take notes during the talk and only after the talk is finished do they see the items to which they have to respond' Hughes 2004:168). The 'complete the notes' task in the Estonian national examination is

equivalent to a gap-filling task – the students have a gapped summary of the monologue and have to fill in the gaps while listening.

Ordering only features once as a task-type. The reason for this probably lies in the complexity of rating such tasks: ‘ if a candidate puts on element of the text out of sequence, it will cause others to be displaced and require complex decision making on the part of the scorer’ (ibid 148).

It is interesting to note that, out of 39 tasks included in the national examination listening section, 24 have to do with completing either the table, the text or the notes (cf. table above), meaning that the vast majority of tasks requires the students to engage in just one type of activity to demonstrate their listening ability – listening for factual information. Other operations that comprise the listening skill, of which for example Hughes (2004) mentions 33 (cf. 161–167), are relatively under-tested. As with the evaluation of writing, the predilection of test writers to opt for only particular task types is bound to have a backwash effect on the classroom practices, resulting in developing the students’ listening ability in a fairly stilted way.

A huge and persistent challenge with the listening comprehension test is quality control of the recordings – finding suitable non-copyrighted texts, choosing speakers for the original recordings (the accent, the speed, the tone of voice, etc. of the speakers), making decisions about the background noise (cf. NE 1997–2008). The latter, however, is an indicator that distinguishes between a B1 level learner and a B2 level learner (CEFR 2001:75) and therefore needs to feature in at least some of the listening tasks intended to identify B2 level students.

2. 5. 3. Reading

Table 6. Task types in national examination reading paper from 1997 to 2009.

Year	Task Type
1997	1. Find summary sentence 2. Match title to text 3. True/false/ no information (TFN)
1998	1. Ordering paragraphs 2. TFN 3.1 Match summary sentence to paragraphs 3.2 Match words and definitions
1999	1. Match questions and answers 2.1 Match summary sentence to paragraphs 2.2 Complete summary close 1.1 Insert sentences in the text 1.2 Match words and definitions
2000	1.1 Match headings and texts 1.2 TFN 1.1 Match headings and text 1.2 Match words and definitions 3. Insert sentences in the text

2001	1. Match questions and answers 1.1 Match texts to questions 1.2 Match words and definitions 1.1 Match summary to text 1.2 Summary close
2002	1. Match title and text 1.1 Insert sentences in the text 1.2 Mach summary sentence and text 1.1 MC 1.2 Match words and definitions
2003	1.1 TFN 1.2 Match words and definitions 2. Match statements and extracts 1.1 Match questions and answers 1.2 MC
2004	1.1 Insert sentences into text 1.2 MC 2. Match questions and extracts 1.1 TFN 1.2 Match words and definitions
2005	1. Match summary sentence and text 2. MC 1.1 TFN 1.2 Match words and definitions
2006	1. Match statements and text 2. TFN 1.1 Insert sentences in text 1.2 Match words and definitions
2007	1. TFN 2.1 MC 2.2 Match words and definitions 3. Insert sentences in text
2008	1.1 MC 1.2 Match words and definitions 2. TFN 3. Match questions and answers
2009	1. Match questions and extracts 2. TFN 3.1 Insert phrases to text 3.2 Match words and definitions

The **reading** paper, similarly to the other papers, derives its topics from the national curriculum. The national curriculum specifies that the reading competencies of a gymnasium graduate include the following: being able to read functionally different texts, including various kinds of instructions; being able to identify both the expressed and the implied main idea of the text; being able to find the necessary or interesting information in the text; can resort to titles, illustrations, drawing, diagrams and fonts to understand the test; being able to guess the meaning of unknown words from context and given elements; knowing the meaning of

more frequent international words in his/her native language and being able to utilise this knowledge when reading a foreign language text; being able to find, choose and use information from various sources in the foreign language; can make use of dictionaries and reference books. (https://www.riigiteataja.ee/ert/act.jsp?id=12888846 29.06.2009) The text types are specified in Year 12 Handbook (specifications) and list 12 different types: 'letters, forms and questionnaires, brochures and prospectuses, posters and leaflets, advertising material, sets of instructions, public signs and notices, menus and tickets, descriptive and imaginative writing, texts accompanied by graphs, diagrams, timetables, and other kinds of non-textual information, informative writing from popular science and reference books, dictionaries (monolingual and bilingual)' (Jõul et al 2005:64). Each reading paper in the national examination has always comprised three texts with hopefully ascending levels of difficulty. What can be noticed looking at the table above that summarises the number and kinds of tasks that follow the texts is, firstly, the markedly varying number of tasks in different papers. While in 1997, each text was followed but with one task, the number of tasks rose to as many as five between 1999 and 2004, to settle at four tasks in the 2005–2009 papers. This may have had an impact on the students' results, as within the given time frame some students had to complete more tasks to get the required number of points than their peers before or after them. Consequently, comparing the results of the test from one year to the next may be problematic at times. The national examination specifications list five task types that should be used to evaluate the students' reading ability: questions (TFN, MC, short answer questions), matching, ordering, completing and information transfer (ibid 65). Looking at the national examination reading papers, it can be seen that the test designers have confined themselves to overwhelmingly to various types of matching tasks – off all the 57 tasks included in the reading papers over the years, 30 are matching tasks. Another widely employed task type is true/false/no information tasks (employed 10 times), with insert the sentence in the text tasks (used 7 times) and multiple choice tasks (6 times) also being popular choices. What we see again is a fairly limited number of methods used for proficiency evaluation, and certain methods being either not used altogether (e.g. information transfer) or being used quite seldom (e.g. various forms of cloze tests). Alderson et al (1995) warn against what is known in the testing theory as the method effect, saying that 'the method used for testing a language ability may itself affect the student's score [….] its influence should be reduced as much as possible.' (1995:44). This means that the student's score can be affected by what type of task he/she is doing. Consequently, by asking the students to complete just particular tasks we may be disadvantaging those who could potentially demonstrate a better level of language proficiency by completing a different type of task. Hence the whole spectrum of task types should be available to students to maximise the potential of excelling at the test for the students. As Alderson et al (1995) assert, 'the more different methods a test employs, the more confidence we can have that the test is not biased toward one particular method or to one particular sort of

learner (1995:45). The task type that has caused perennial debate within the paper is the true/false/no information task that places huge demands on the item writers to create items which clearly belong in just one of the given categories (true or false or no information) and is not interpretable in more than one way (see for example http://www.opleht.ee/Arhiiv/2007/25.05.07/aine/3.shtml 30.06.2009). Research literature seems to have a generally cautious attitude to that particular task type, Hughes (2004) claiming that 'there is no place for items of this kind in a formal test' (2004:79) and Alderson et al (1995) affirming that '[T/F/N] tasks can be useful in a reading comprehension test, but […] it can be demanding and lead to student confusion' (1995:51). Attempts could be made to replace this task type with as effective but a less problematic task type.

2. 5. 4. Language Structures

Table 7. Task types in national examination language structures paper from 1997 to 2009.

Year	Task Type
1997	1. word formation 2. insert a word where necessary 3. banked gap-filling 4. insert a correct verb form
1998	1. banked gap-filling 2. insert a correct verb form 3. word formation
1999	1. insert correct articles 2. insert correct prepositions 3. insert a verb form where necessary 4. use the appropriate verb form
2000	1. insert the correct tense 2. insert a word where necessary 3. insert a correct verb form 4. insert a correct verb form
2001	1. a banked gap-filling – degrees of comparison 2. create a sentence by putting the words in the correct order 3. a banked gap-filling – pronouns 4. insert a word where necessary
2002	1. insert a correct verb 2. insert a correct article or a preposition 3. insert a correct question tag 4. word formation
2003	1. insert a correct article or a preposition 2. MC 3. banked gap-filling
2004	1. paraphrase – indirect speech 2. word formation 3. insert a correct article or a preposition 4. insert a correct verb-form

2005	1. insert correct prepositions 2. paraphrase using modals 3. paraphrase using a correct verb form 4. word formation 5. create a sentence by putting the words in the correct order
2006	1. delete the unnecessary word 2. MC 3. word formation 4. MC
2007	1. banked gap-filling 2. MC 3. delete the unnecessary word 4. word formation
2008	1. MC 2. delete the unnecessary word 3. insert the correct word 4. banked gap-filling
2009	1. a banked gap-filling 2. MC 3. word formation 4. delete the unnecessary word

The **language structures'** paper focuses most specifically on the grammatical accuracy and appropriacy of the English language use. It is this part of the language competence that has been specified in the most detail in the national curriculum (for the list of grammatical requirements for a upper-secondary school/gymnasium/high school graduate see, for example, Jõul et al 2005, appendix E, 131–133). The challenge for the test writers has been to achieve appropriate coverage of the specifications. If well designed, this section allows "checking the students' knowledge within a fairly short amount of time of very different language structures, also those that in a daily language feature less frequently" (NE 2001:19). The Year 12 Handbook (2005) also specifies the task types that are to be used during the evaluation: gap-filling, banked cloze, multiple choice, ordering words, paraphrasing, word formation in sentence/ passage completion, word formation and position of a word in a passage, editing (deleting an irrelevant word from the text/ line/ passage, combining clauses and sentences, various combinations of task types listed above (84). The table above shows the task types employed in different papers. The number of tasks has changed very little: only in 1998 and 2003 were the students asked to complete 3 tasks and in 2005 – five tasks, all the other papers comprise four language structures tasks. As with previous papers, it can be noted that the test developers seem to prefer particular task types to others. Gap-filling of various types predominates in all the test papers to the exclusion of almost everything else. Editing, multiple choice and word formation have been utilised slightly more frequently since 2006. At the same time, paraphrasing and combining clauses and sentences are tasks not often resorted to. There has been a change in the way grammar structures are tested. While in the great majority of cases grammar structures are checked within complete, connected texts, there was a period (2003–2005), where the tasks consisted

of isolated sentences checking a particular discreet item. For students on this level, it is not sufficient to be familiar with particular grammatical items only, it is necessary to know how to implement the grammatical knowledge within a particular text. B2 level students need to operate on the text, not on the sentence level (cf. CEFR 2001:61), so completing tasks like this may not allow them to show their full proficiency. A successful completion of tasks requires attentive reading of the tasks on top of grammar knowledge. It is here that we notice that dividing language tests into skill tests is somewhat arbitrary in that by testing one skill we are inadvertently also testing another (in this case, while testing structures, we are also testing the reading skill).

2. 5. 5. Speaking

Table 8. Task types in national examination speaking section from 1997 to 2009.

Year	Task Type
1997	1. Picture description and discussion 2. Role-play
1998	1. Picture description and discussion 2. Role-play
1999	1. Picture description and discussion 2. Role-play
2000	1. Picture description and discussion 2. Role-play
2001	1. Comment on a quotation. 2. Role-play
2002	1. Comment on a quotation. 2. Role-play
2003	1. Summarise a reading passage and comment 2. Role-play
2004	1. Summarise a reading passage and comment 2. Role-play
2005	1. Summarise a reading passage and comment 2. Role-play
2006	1. Summarise a reading passage and comment 2. Role-play
2007	1. Summarise a reading passage and comment 2. Role-play
2008	1. Monologue based on a common belief 2. Role-play
2009	1. Monologue based on a common belief 2. Role-play

The **speaking** test takes place on a day following the written papers (depending on the size of the school, it may take between 1 and 3 days to administer the speaking test to all the students who have registered for it) and currently requires the examinee

to complete two tasks: a monologue and a (two-participant) role-play. The interview is administered by an interviewer and assessed by an assessor who has not been the candidate's language teacher. The interviewer's function is to make sure a consistent procedure is applied from one candidate to the next, but he/she is not to participate in the evaluation of the candidate. The assessor does not participate in the interview or interact with the candidate. His/her only function is to evaluate the student's performance.

The national curriculum specifies that the speaking competencies of the a gymnasium graduate include the following: using the intonation, rhythm and emphasis characteristic of the foreign language; can converse within the topics given in the curriculum, present and account for one's points of view; being aware of and using the etiquette of interaction; being able to interact in the foreign language both directly and via the telephone; being able to exchange information, ask questions and express opinions on social problems and events; being able to resort to compensation strategies when necessary (https://www.riigiteataja.ee/ert/act.jsp?id=12888846 29.06.2009).

The prompt for the monologue has gone through a thorough process of evolution, proceeding from a picture (until 2001), to a quote (2001–2002), a short text (2003–2007), and as of 2008, a controversial statement. The main reason for the most recent change, substituting short articles as prompts, was the attempt to reduce the amount of reading in the speaking test. As can be seen from the discussion above, the national examination already has a fairly heavy bias on testing reading (the reading paper, and the language structures' paper). The new format allowed the examinee to focus on displaying his/her speaking skills without depending on the reading-comprehension first. This part of the national exam has been updated most recently for the purposes of higher reliability. Both tasks of the exam are scripted, i.e. the interviewer has to follow a prescribed format for the interview and is not allowed to improvise or deviate from the wording of the script. Improvisation may lead him/her to ask questions of varying levels of difficulty from different examinees, leading to unequal treatment and potentially unfair marking. Following a script will ensure equal conditions for all examinees, irrespective of the examination day, the time of the day, the order of the examinees and the fatigue level or the personal characteristics of the interviewer.

The speaking section of the national examination paper, though having gone a process of validating its format, assuring consistency of the procedure from one interview to the next, making sure that the interviewers and assessors have undergone the relevant training and that the assessors know how to implement the marking scale to the performance, still has the underlying problem of having very low validity. The problem lies in the absence assessor monitoring system, of establishing inter-rater and intra-rater reliability. The reason for this is that the interviews are generally not recorded, or are recorded only in case the student requires it. Owing to this, there is little information about what actually happens during the oral interviews, how consistent the interviewers are

during the interview and how consistent the assessors are in their marking. This is one of the problems that the current research hopes to establish during the subsequent research.

2. 6. MARKING PROCEDURES

Both objective and subjective marking have been implemented with the national examination in the English language from the very start. Listening, reading and language structures' papers have always been marked objectively, relying on the answer key for each item. Providing the answer key is a simultaneous process to item writing but also continues during the piloting stage, which invariably produces occasional acceptable but previously overlooked answers. Once the answer key is complete, no judgement is required on the part of the marker. A special case are the tasks in the listening paper that require students to fill gaps or provide short answers, and consequently issues of correct spelling come into play. Here, a complete answer key cannot be prepared prior to test administration. To ensure uniform marking, a standardisation meeting is usually called after the examination paper has been administered and a random sample of about one hundred papers is taken to determine the extent of spelling diversion accepted as correct. In principle, no 'points for errors of grammar or spelling [are deducted], provided that it is clear that the correct response was intended' (Hughes 2003:170). It is, however necessary to determine where the line of clarity runs. When the respective decisions are made, the marking proper will proceed according to the key compiled.

Writing and speaking sections of the national exam are subjectively marked, i.e. teams of raters are trained either to rate the students' writing papers or their performance during the speaking test. In writing, the raters have generally relied on two different marking scales – one for letters and another for the essays and reports. With the number of points available for a particular paper fixed – 20 points as a sum total for both tasks – the major concern while developing the marking scales has always been what to reward within the skill. The marking scale for letters has moved from awarding points for task completion, letter format and language (1997) to evaluating task completion, vocabulary and register, and grammar and spelling (2001), to task completion, letter format and language (until 2006) and task completion and language (as of 2007). It is also interesting to note that until the 2007 scale, specific sub-skills had been weighted differently. An example is the 1999 scale, where for task completion the students could get the maximum of 2 points, but for vocabulary and register and for grammar and spelling a maximum of 3 points. In the 2006 letter scale, task completion and format both earned the writer a maximum of two points, but the language criterion was evaluated on the scale of 0 to 4. This type of marking may inadvertently lay the classroom teaching emphasis on language (i.e. grammar and vocabulary) and overlook other facets of writing, such as content and organisation, thus disadvantaging the student, should he/ she move to such language contexts where

the aforementioned qualities of writing are required. For a more detailed discussion of the 2007 national examination writing scales see Alas et al 2006. All writing papers are marked by two raters and in case of a disagreement of 4 points or more in the evaluation results, a third rater is called in for a final decision.

The marking of speaking has undergone substantial changes, too. The challenges for the rating scale development are similar to those with the writing scales, i.e. which criteria to select for evaluation. Here, too, the scale has moved from a full scale for all the criteria selected in 1997, to an unequal number of points allocated for different criteria (as of 2001) back to a full scale starting from 2007. The current marking scale evaluates the students' performance from the point of view of four criteria − communication, vocabulary, grammar, and pronunciation and fluency. For a full discussion of the 2007 speaking scale see Alas 2007. The students' oral performance is rated by an independent assessor during the oral exam. The assessor does not participate in the interview, which takes place between the student and the interviewer, but only rates the student's performance relying on the marking scale.

2. 7. EXAMINATION RESULTS

All five exam sections are equally weighted – the maximum number of points that can be awarded for each section is 20, thus the maximum number of points the examinee can receive for the whole exam is one hundred. Below, an attempt will be made to draw some conclusions from a decade of the English language national examination administration in Estonia. The comparison and analysis will rely on the national examination 1997–2007 results. The table below shows the average scores of the students who have taken the national exam in the English language over the years along with the standard deviation i.e. the 'average amount that each student's score deviates from the mean' (Alderson et al 1995), the maximum number of points gained and the minimum scored during a particular test.

Table 9. Examinees and their mean score.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Examinees	9280	8769	9258	9461	8488	9311	9431	9099	9415	9590	9696
Average	64.6	58.8	61.8	64.1	64.9	66.6	63.99	66.6	71.9	64,4	68.8
Std*	17.7	19.9	19.9	19.7	18.8	17.8	16.9	16.7	16.0	16.1	16.0
Max	99	99	100	99	99	100	100	100	100	99	99
Minimum	8	0	0	0	5	0	0	1	1	11	5

* std = standard deviation

Looking at the average scores, which is just one of the very many statistical data derived from each year's test result, it can be observed that with two exceptions the mean score has remained relatively stable during the decade. It is only in 1998, that the average score has dropped to 58.8 points, which may indicate a relatively

more difficult test compared to the others. In 2005, however, the average score suddenly shoots to 71.9, which in turn points at a somewhat easier national exam. With these two exceptions, the examination development team has managed to produce fairly uniform exams.

It is also worthwhile comparing the average scores awarded for particular skills within the exams. The table below makes comparisons between the average scores calculated over the years (1998–2007) for a particular skill as well as juxtaposes it with the averages for the other 4 sections of the test.

Table 10. Overview of mean scores for skills (1998–2007).

Year	Writing	Listening	Reading	Structures	Speaking
1998	12.2	10.1	10.7	10.4	15.6
1999	12.4	11.2	10.9	11.8	15.7
2000	12.3	11.6	13.3	9.9	15.6
2001	11.3	14.7	12.2	11.1	14.7
2002	11.6	13.2	14.7	11.9	15.5
2003	11.5	11.9	13.5	11.0	15.8
2004	13.4	12.0	13.7	11.5	16.1
2005	13.3	12.7	15.3	13.1	16.4
2006	12.9	11.3	11.9	12.1	16.6
2007	13.1	13.1	12.5	13.1	16.9

Comparing the results across the board, it can be seen that while writing, listening, reading and language structures seem to correlate fairly well with one another, the average score for speaking is significantly higher every year. If these scores are reliable, then the students' speaking skills are for some reason significantly higher than all the other skills. Given that successful speaking presupposes good vocabulary, a good command of grammatical structures and the ability to interact with the interlocutor (hearing, understanding and responding to what is said, i.e. listening skills), the result is somewhat dubious from the point of reliability. Another factor that may skew the results is the fact that although the schools are urged to record the examinees, and the examinees are urged to request recording of their oral interviews (without a recording the student cannot appeal against their interview result), this is not general practice. Thus all the interviews are marked by just one rater whose judgement is hardly ever monitored, which may lead to a tendency to inflate the score in an attempt to compensate for possible lower scores in other sections of the test.

The students' average results have already been discussed above. It would, however be interesting to look at different groups of students. The table below shows the average results of male and female students from the time when such comparative data are available.

Table 11. Mean score of boys and girls.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007
Boys	60.3	61.0	63.3	66.2	63.3	65.5	71.4	65.2	69.5
Girls	62.8	63.6	64.6	66.9	64.4	67.3	72.3	63.8	68.3

The table shows that with two exceptions (2006 and 2007), the girls' results have generally been higher, which may indicate a slightly better language competence level of girls, but could also be an indicator that the exam items have been constructed so that they are more accessible to the female population of test takers. From the raters' comments it seems to transpire that girls are generally better at completing writing and speaking tasks while boys are more successful in listening, reading and language structures.

Another point of comparison is the medium of instruction at school. Estonia has both Estonian and Russian language schools, where the primary language of instruction is Estonian or Russian, respectively. The same exam is available as a national exam for both school types. The average results of the students can be seen in the table below.

Table 12. Mean score of Russian and Estonian students.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Estonian	66.8	61.2	64.4	64.6	65.8	68.3	65.6	67.5	74.1	66.3	70.7
Russian	59.2	51.5	53.5	55.6	59.1	61.8	59.3	64.2	65.2	57.8	68.8

A study of the results demonstrates a significantly higher average every year of the students studying in the Estonian language schools. The difference may be explained by the fact that while most of the Estonian-speaking test takers have studied English as an A-language (the first foreign language that the students start studying), the vast majority of the Russian-speaking students taking the test have started studying English as a B-language (the second foreign language, which begins two years after they started their first foreign language – Estonian). Thus by the time the examinees take the exam, the Russian students would have studied English for a shorter period of time.

2. 8. CONCLUSION

The historical background to the examination has been provided to demonstrate the development of the construct of English language proficiency evaluation as it is today, to understand how the current framework for the examination was reached. It was also considered important to demonstrate the effort that was put into creating a valid measurement instrument, and to show the learning curve in the field – language proficiency evaluation as it was understood in Europe and America at the time – that

was virtually non-existent before 1994 in Estonia. Examination validation as was shown in chapter one is all about providing evidence, needs to be about content relevance, criterion relatedness and meaningfulness of construct (Bachman 1990:243). Tracing the steps that were taken in the exam development serves, I hope, as providing that evidence.

The above discussion allows us to draw a number of conclusions. A thorough preparation period preceded the launch of national examinations in Estonia, which has resulted in the national examination in the English language today being based on a construct that interprets language competence as consisting of four skills yet deems that control of language structures also be part of language competence measurement.

The English language national exam is well established. It is the most widely taken, locally constructed, nation-wide foreign language proficiency exam in Estonia. The examination design proceeds from the national curriculum and the test specifications elaborated in Year 12 Handbooks. The Estonian national curriculum specifies B2 as the language level required in English from the Estonian gymnasium graduates. The curriculum outlines in very broad terms the different CEFR levels but to date, the levels have not been sufficiently elaborated in the curriculum. The national exam in English is a B2 level exam insofar as it proceeds from the CEFR framework and tries to align its tasks and language content with other English language proficiency exams that have the B2 status (e.g. FCE).

Designing tasks for the national examinations, the writers seem to be aware of the need to vary the measurement instrument in terms of task types, yet on closer observation the designers seem to be confined to particular task types more than others, which may affect the overall test reliability and validity.

Compared to other skills, speaking is the most vulnerable to validity queries, as the candidate performance is not subjected to systematic second marking. Neither can the results of the first marking be validated by random remarking.

3. THE CURRENT FRAMEWORK FOR TESTING ORAL PROFICIENCY AT THE NATIONAL EXAMINATION OF THE ENGLISH LANGUAGE IN ESTONIA. A QUESTIONNAIRE

The current chapter will commence to view the validation process of the speaking section of the national examination in the English language in Estonia. In order to do so, it will first give a rationale for the current speaking section of the national examination and the framework that has been adopted for the section as of spring of 2008. Then, it will discuss a questionnaire study carried out among the teachers who implemented the new framework during the national examination.

3. 1. A NEED FOR A MORE RELIABLE EVALUATION INSTRUMENT

While developing valid tests to evaluate oral proficiency, a test designer needs to be aware of the essence of oral language ability, which Hughes has defined as 'to interact successfully in that language...(which) involves comprehension as well as production' (Hughes 2004:113). For the actual test construction we thus need to first, find out the repertoire of oral tasks the candidates should be able to complete in the target language, and then make a choice to formulate the tasks for our particular exam situation that would 'elicit behaviour which truly represents the candidates' ability' (ibid).

Luoma (2004) describes the assessment of speaking as a cycle of five stages, which starts with a realisation of a need for a candidate's speaking score. This is followed by a stage where an instrument is designed that can be used to obtain the score needed (a test development procedure where three essential elements are devised, tried and revised: tasks to elicit a rateable sample of the candidate's language, a rating system to evaluate the candidates' performance; and quality control measures). The third stage is the test administration/performance where the candidates demonstrate their speaking ability and which is often recorded on a video- or an audiotape. This is followed by the process of rating/evaluation where the raters apply the rating criteria to the candidates' performance to produce a score. The cycle ends with the use of the obtained scores to meet the needs that prompted the development of the test in the first place (Luoma 2004:4–5). Luoma emphasises the overarching importance of quality assurance during the whole cycle. 'The main qualities that the developers need to work on are construct validity and reliability' (Luoma 2004:7).

Construct validity being our main concern, the test developers need to first identify what has been specified as construct. The construct for the national examination speaking section has been specified in the national curriculum. According to that, a student demonstrates oral proficiency through 'employing the correct foreign language intonation, rhythm and stress; being able to converse within the specified topical

range by presenting and supporting their point of view; by knowing the communication etiquette and being able to use it; by being able to communicate in the foreign language both directly and by telephone; by being able to exchange information, ask questions and express their position on social problems and events; and by resorting to compensatory strategies in communication if necessary' (https://www.riigiteataja.ee). The topic areas specified in the national curriculum are the following: I as an individual among other individuals, my special features, abilities, preferences, strengths and weaknesses; family and home, marriage and family, roles in the family, rights and obligations, family budget; friends, relations between friends, social problems; environment, Estonia, the world, nature and nature protection, natural resources, climate, town and country, urbanisation, Estonian government, economy, cultural traditions, international relations; English-speaking countries, governments, culture, international relations; everyday activities, healthy ways of life, nutrition, communication in service situations, help during emergencies; study and work, the system of education, opportunities for education in Estonia and English-speaking countries, study skills and exam techniques, work and unemployment, technological advancement; hobbies and culture, sports events, cultural figures, advertising, information society and its problems (ibid).

The discussion in the previous chapter revealed that the national examination specifications are not only drawn from the national curriculum but also from the document Year 12 Examinations, which stipulates the test focus in further detail: 'students should be able to use appropriate conversational formulae for greeting, leave-taking, etc.; indicate likes, dislikes and preferences; express agreement and disagreement; express opinions; give instructions; ask for and offer help/ advice; ask for factual information; respond to requests for factual information; state basic conceptual meanings, e.g. numbers, times, dates, quantity, location; make arrangements concerning time and location; describe objects, individuals and sequences of events; maintain a conversation on a variety of topics; deal with a breakdown in communication by indicating lack of comprehension and asking for and offering clarification (Jõul et al 2005:97). The same document gives a list of task types that can potentially be used to elicit the aforementioned subskills: describing a picture, telling a picture story, describing charts graphs, tables, diagrams; asking/ answering questions, participating in an information gap activity/discussion/ conversation/role-play; giving a short unprepared monologue (comment) on a given topic (ibid).

As there is no theoretical discussion offered in either document in terms of what constitutes the construct of the speaking ability, conclusions have to be drawn from the kind of information given in them discussed above. What can be observed from the data above is that the construct of the spoken language seems to be perceived as being composed of at least three elements: the linguistic element, which comprises the knowledge of lexis on the given topics, as well as phonological knowledge of rhythm, stress and intonation, and syntactic knowledge of the grammatical categories and structures listed; the sociolinguistic element, which refers to the candidates' ability to select an appropriate politeness level and being aware of the differences between

Estonian and British/American cultural norms; the pragmatic element, which covers the language functions, the candidates' ability to implement particular interactional scripts that they have internalised. This understanding of the essence of the spoken language corresponds to the theory set forward by CEFR (2001:13) and allows us to speculate that on the basis of the specifications given in the national curriculum and in the Year 12 Handbook a test can be developed that could potentially be aligned with the recommendations of the CEFR in the future.

Proceeding from the requirements set above, i.e. to give the students a chance to demonstrate both comprehension of oral language and its production, and do that within the topic range specified, the national examination oral section has been put together to comprise three sections (introduction, task 1 and task 2), two of which (task 1 and task 2) are rated and the initial one (introduction) serves as a lead-in and a warm-up and is not rated.

The first rated task before 2008 was a monologue that derived its topic from a short article a student had to read before his/her monologue. The student was expected to summarise the article without retelling it and comment on the problem it posed. The monologue was followed by a set of questions from the interviewer which were designed to expand the discussion initiated by the student. The second task was a role- play between the student and the interviewer, the completion of which required the command of a slightly different discourse.

It is inevitable that language testing tools, in our case national examination oral proficiency testing techniques, are constantly reviewed in keeping with the increase in our knowledge about language teaching and language testing. One of the main considerations in language testing is that 'it is important that the techniques used should interfere as little as possible with the (skill) itself, and they should not add a significantly difficult task on top of [the skill tested] (Hughes 143). Unfortunately, so far this seemed to be the case in the speaking section of the national examination, more particularly in the first task, where, in order to perform a monologue, the students had to spend 15 minutes familiarising themselves with a short reading text. Although integrated tasks, where a student is, for example, first asked to read a text and then to listen to an audio-text on a related topic and finally comment on what he/she has been reading and listening, are not uncommon in language testing (Luoma 43), this approach was not considered suitable in this particular section, because two skills seemed to be explicitly involved (reading and writing), but only one of them was being assessed (speaking). Also, all the other sections (writing, listening, reading and language structures) had also adopted a construct-based approach as far as possible, so it seemed only appropriate that the same approach should be adopted here. There were other considerations that motivated the change of the prompt. A reading prompt was seen as testing once again the students' reading ability that had already been tested on a previous day in quite an extensive reading section (three reading extracts of approximately 1000 words altogether followed by 4 reading comprehension tasks). Furthermore, the reading texts chosen as prompts for the speaking examination, besides requiring a substantial effort to find (at least 60 concise, topically appropriate,

equally difficult texts were necessary every year), varied considerably in the level of difficulty and sometimes did not include an easily discernable problem. All this put the students at an unequal position while preparing for their monologue, creating reliability issues right from the start. The test development problem that was thus posed was to find a way of setting a prompt to the students that would remove the task of reading but that would, nevertheless, reflect the test specifications as much as possible and elicit a rateable sample of spoken language from the candidate.

The second task of the speaking section – a role-play between a candidate and an interviewer was retained in its earlier form as a task-type that would elicit language that would not necessarily emerge during the monologue, e.g. seeking information, seeking for clarification and responding to requests for clarification, complaining, apologising, complimenting, etc. Furthermore, it is during the role-play that the candidates' control of particular sociolinguistic and pragmatic language features comes to the fore. The task itself was thus considered appropriate; there were other aspects surrounding the task that needed adjustment in the test developer's view, which will be discussed below.

Referring back to the testing cycle above, it was not only the task content that determined how reliable and consequently valid speaking assessment was going to be. Part of overall reliability is rater-reliability – how trustworthy the work of the interviewers and raters is during the test. Alderson et al (1995) sum it up best saying that 'whether interlocutor or examiner, the person who interacts with the candidate must be in control of techniques which will help each individual to feel at ease, while at the same time paying attention to details such as timing and wording of prompts, to ensure that all candidates have an equal opportunity to display their abilities' (1995:116). The speaking section of the national examination of the English language has been standardised in very general terms only. Although, the instructions have generally stipulated that the oral interview should last about 15 minutes, there are loopholes in the instructions that lead the research to conclude that the time spent on individual interviews may vary considerably, which makes comparison of the individual interviews problematic and consequently undermines the reliability and validity of the results of the speaking section. For example, the time allotted for each particular interview according to the instructions, may fluctuate between 10 and 14 minutes, resulting in some students taking almost a third of the time longer to be assessed (NE 2001:84–45). Also, the guidelines specify that the student can spend 5 minutes (which are not considered when measuring the overall interview time) familiarising themselves with the text, yet proceed to instruct that the student will continue with the monologue when 'they are ready to start' (ibid), which may mean a considerably longer preparation period, i.e. unequal conditions. The same guidelines allow the interviewer to ask additional questions once the candidate has finished his/he monologue. The guidelines give the impression that asking additional questions is optional on the part of the interviewer and thus create a situation where no questions are asked where additional questions could perhaps have enhanced the candidate's performance (and the interviewer did not ask the questions because he/she could not

think of anything to ask or was too tired to ask the questions or thought perhaps that the candidate would not be able to respond to them). Alternatively, the interviewer could ask an infinite number of questions (although the candidate has already given an exhaustive response to the prompt) just because the interviewer was carried away by the topic or the response the student has given, resulting in time being spent on obtaining a sample that has already been amply obtained. In both cases, the students will be in unequal positions while being interviewed compared to their peers. Even if the cases are not as extreme as described above the number and the kinds of questions the interviewer asks, if not pre-tested prior to the interview proper, may result in a considerable change in task difficulty. Consequently, another task was to create a set of tools for the interviewers that would make sure that all the candidates are addressed in a standardised manner during the interview. The tools, if properly used, would considerably unify and clarify the interviewer's conduct during the interview on the one hand and ensure a unified treatment for all the candidates, on the other, thus adding to the overall reliability and validity of the testing of speaking procedure.

In addition to standardising the interviewer behaviour and language, a key component in the spoken language proficiency evaluation process is the marking scale. The previous marking scale had been introduced in 2001 and although the speaking prompt was changed in 2003, the marking scale was retained without change (cf. NE 2001, NE 2003). With the changes in the tasks, alterations in the marking scale seemed inevitable. All the marking scales for subjectively marked sections of the national examination have been developed relying on the intuitive methods, i.e. 'on the principled interpretation of experience [where] the developers may consult existing scales, curriculum documents, teaching materials and other relevant source materials and then distil the information into draft descriptors at an agreed number of levels' (Luoma 2004:83). The draft versions are then discussed, piloted and edited by the test development team and further minor adjustments can be made once the scale has been implemented by all the assessors on the basis of the feedback they give. The 2001 marking scale for speaking was considered problematic for mainly two reasons, the inconsistency of the criteria in the scale as well as the weighting of the particular criteria in the scale. The criteria included in the marking scale are monologue, pronunciation, vocabulary, grammar and communicative ability where the first element (monologue) is quite clearly a task type the completion of which requires utilisation of grammar, vocabulary and communicative ability with an acceptable level of pronunciation. What the assessor seemed to be required to evaluate under the criterion monologue was the time that the candidate keeps going and the structure of the monologue. It is not clear if the other criteria are to be considered while evaluating the monologue. There is also overlap in the descriptors in that being prompted or not during the interview gets assesses within the monologue as well as under communicative ability. A further inconsistency is within the criterion of pronunciation, where the presence of accent may yield the maximum 3 points, but may also be the reason for affording just one point (cf. NE 2001). The other consideration was the weighting of particular criteria. As can be seen in the 2001

scale, it is the vocabulary and grammar that earn the greatest number of points during the speaking test. The students earn fewer points by demonstrating excellent communicative ability and still fewer points by demonstrating excellent pronunciation. Brown and Hudson (2002) warn against creating a negative washback effect if there is a mismatch between the curriculum objectives and the test (2002:48), which seems to come to the fore here. As can be seen in the discussion above concerning the test specifications, both phonological and communication aspects are specially emphasised in the national curriculum. The marking scale, however, seems to refute the need to develop those qualities. By affording fewer points to particular aspects of students' speaking performance (in this case phonology and communication) will result in those aspects being less taught in the language classrooms to the detriment of the students' general speaking ability.

To conclude, in order to design a more reliable testing instrument for measuring the candidates' speaking ability, alterations seemed necessary in the task, the test administration and the scoring system of the candidates' performance.

3. 2. THE FRAMEWORK FOR THE SPEAKING SECTION OF THE NATIONAL EXAMINATION 2008

The new framework for the speaking section of the national examination in the English language introduced in 2008 included the preparation of the following documents:

- Scripts for the three stages of the speaking section
- A new marking scale for speaking
- Guidelines for the interviewers
- Guidelines for the assessors
- A training course for the interviewers and assessors

This is in accordance with the task-related documents and materials list proposed by Luoma (2004), which deems the following elements necessary to develop a speaking assessment instrument: 'the rubric and the instructions to examinees; the task materials, which the examinees use while performing the tasks (if relevant); an interaction outline, which gives guidelines or scripts for examiners about the content and wording of questions or prompts; plans and instructions for administration' (2004:51).

3. 2. 1. Scripts for the Three Stages of the Speaking Section

The structure of the speaking section was retained: an introduction, followed by the monologue and follow-up questions (task1), followed by the role-play (task 2). The overall time was estimated by samples drawn from other international examinations (FCE, CPE, IELTS) as well as the experience obtained over the years in

Estonia while testing speaking within the framework of the national examination and was initially aimed at not exceeding a 15- minute time-limit. The concrete slots allocated for each part of the interview were finalised during pilot testing and were settled as follows: the introduction should not exceed 2 minutes; task 1 should take between 8 and 9 minutes and task 2 between 4 and 5 minutes. Thus, the overall time required to get a rateable sample from each candidate would be between 14 and 16 minutes. The fluctuation within each part would not exceed a minute and the overall difference in the time taken to interview individual candidates should not vary beyond two minutes. The variation should not derive from the difference in number of questions asked of each candidate but rather from the candidates' or interviewers' rates of speaking. From the point of validation, maintaining the time-frame with all the candidates examined was considered one of the first steps towards a more reliable evaluation system. An implication of this decision was to ensure that the examination environment was equipped to meet that requirement, which in its turn meant training interviewers to keep time during the interview and providing examination rooms that would allow unobtrusive time-keeping.

In order to further standardise the procedure, the notion of scripts was introduced. Scripts are scenarios that the examiner uses verbatim while guiding the student through the examination procedure. Their aim is to standardise the examiner language and behaviour and consequently increase the level of standardisation of procedure. The decision to opt for scripts rather than guidelines stemmed from the experience gained during in-service teacher training, where numerous questions were asked about the precise wording of particular sections of the interview. Given that oral interviews during the national examination in the English language are conducted (with very few exceptions) by non-native English speakers in Estonia, scripts were considered not only more helpful but also infinitely more reliable.

Three scripts were created – one for each stage in the examination – the introduction, stage 1 (monologue) and stage 2 (role-play).

A prototype script for stage one is given below.

STAGE 1: Introduction (maximum 2 minutes)

Greet the candidate and ask him/her to sit down.
Ask the external candidates **if they are familiar with the procedure** / explain if necessary.
Ask the candidate **if he/she wants the interview to be recorded**.
If 'Yes', switch on the cassette recorder and state the candidate's code number.
If 'No', ask if the candidate is aware that he/she can only appeal against the result of the speaking paper if the answer is recorded.

Interviewer: **Hello. (**If the candidate does not know you, tell him/her your name**) I am your interviewer today, and this is (name), your assessor. How are you today?**
If candidate responds, 'I'm fine', proceed with ***'That's good then.'***
If candidate responds, 'Quite nervous', proceed with '***Just try to relax. You'll be fine'*.**

Choose **ONE** of the following scenarios to continue (vary them equally during the day):
Interviewer: **Let's talk about the weather. Do you like the weather today? Why? What is your favourite type of weather? Why?**
Thank you.

OR

Interviewer: **Let's talk about your home. Do you live in a house or a flat? What do you like about your house/ flat? Why?**
Thank you.

OR

Interviewer: **Let's talk about photographs. Do you like to take photographs? Why? Why do people usually like to look at photographs?**
Thank you.

OR

Interviewer: **Let's talk about computers. Do you like working with a computer? Why? What do people usually use a computer for?**
Thank you

The aim of stage one is to lead the student into the examination situation and to establish a relaxed professional atmosphere. The sets of warm-up questions (i.e. scenarios) are written out verbatim for the interviewer, and the examiner should not improvise here or paraphrase them. Lazarton (1996) has found that 'unequal interlocutor support may well lead to bias in ratings' (qtd. from Reed and Cohen 2001:86) while Reed and Halleck (1997) assert that 'level and type of questions have, for example, been found to influence ratings of the very same candidate when interviewed by different interviewers' (qtd. from Reed and Cohen 2001:86). Thus maintaining the warm-up questions intact from one candidate to the next ought to further sustain reliability. The topics/questions suggested above are samples, the actual warm-up questions will vary on the examination script, they will vary from one examination day to the next and from one year to the next. They will be drawn from the national examination specifications but they will, however, be purposefully easy questions to understand and answer even to the weakest candidates. These questions are only there to warm the student up for the language and give him/her the feeling of being in control, the feeling that the forthcoming examination will hopefully be within his/ her level of competence. The candidates' answers to the warm-up questions are likely to be short and predictable, at no point should the interviewer be tempted to encourage the student to speak long on any of the warm-up questions. The responses given by the candidate during this section of the interview should not be considered while rating his/ her language.

Once the introduction has been completed, the interviewer will resort to the next scrip to lead the student to the first evaluated oral examination task. A prototype script for stage two can be seen below.

STAGE 2: Task 1

Interviewer: ***Now, I would like you to speak on a topic for two minutes. Before you talk, you have 3 minutes to think about what you are going to say. You can make some notes if you wish.***
Do you understand?
Here is a pencil and some paper. [hand over pencil and paper]
Please, pick a topic. [point to the cards on the table]
What's the number of your topic?
Now you have 3 minutes.

The candidate has uninterrupted preparation time for 3 minutes. (The cassette recorder should NOT be switched off for that time)
When the time is up, stop the candidate by '***Alright. Remember, you have two minutes for speaking. I'll tell you when the time is up. Please start speaking now***.'

Allow the candidate **2 minutes** of uninterrupted monologue time.
Sample Topic:
Some people think that physical education should be on students' timetable every day. Why do you think they say that? Do you agree? Give reasons.

When the candidate has been speaking for 2 minutes, find a logical way (at the end of a sentence or thought) to stop the candidate in a natural and friendly manner.

OR

When the candidate has spoken for less than 2 minutes and it is not clear if he/she has finished, ask ***'Is that all you wanted to say?'*** or ***'Was there something else you wanted to say?'***
When the candidate has completed the monologue, continue with the questions in the script in the same order they appear (unless the candidate has already answered any of them in his/her monologue, in which case skip the question).

Interviewer: ***Thank you. Now, I would like to ask you some questions.***
1. What were your favourite subjects at school? Why?
2. How important is sport in your school?
3. Why do people like some subjects more than others?
4. Can schools prepare students for life? Give reasons.

Once the candidate has finished, mark the end of the task by
'*Thank you. Let's move on to the next task*'.

The prompt for the monologue, as can be seen from the script above, is a controversial statement that relates to students' everyday life and surroundings. It is a conviction that candidates should be able to comment on, on the basis of either their life experience or what they have read about or seen in the media. The topic of the statement is drawn from the examination specifications and it is worded so that it should provoke disagreement with it at first sight. The justification for this kind of wording of the prompt is that having to give arguments for a belief first and then counterarguments for an opposite view would generate more language than focusing just on agreement and justification. Wording the prompt as a controversial statement reduces the amount of reading to a minimum (compared to

the earlier prompt) and allows the candidate to start focusing on oral production immediately. An effort is made to maintain age and gender appropriacy (e.g. avoiding a bias towards stereotypically known as 'male' or 'female' topics, or statements that the candidates would find cognitively difficult to comment on) while developing the prompts but more research is needed in terms of how effective it has been so far.

The determination of how much time the candidates should have for both preparation and the monologue per se was determined relying on what the candidates might be expected to do if they take other related examinations, e.g. IELTS or TOEFL, if they want to proceed to study at the universities abroad. Both the above-mentioned examinations allow between one (cf. TOEFL) and two (cf. IELTS) minutes to respond to the monologic prompts with about a minute to prepare for the response. The respective time slots were set at two minutes for the speaking time and 3 minutes for the preparation time. Although the preparation time within the international examinations is considerably shorter, extra time was thought feasible relying on the recommendations of the teachers, the previous national examination experience (which had about 5 minutes for the preparation of the monologue), and the fact that the current national examination has a slightly different purpose compared to the international samples. It is important from the examination validity's point of view to be rigorous about time-keeping in all stages. Consequently, learning the habits of time-keeping as well as proper techniques for stopping students when their time is up and making smooth transitions to the subsequent interview stages should be inseparable parts of interviewer training. Cases in point in the above script are 'alright' to signal that the candidate's preparation time is up, and 'thank you' to mark either that the students has exhausted the time that has been allotted for the monologue or that the whole task has been completed.

Each monologue prompt consists of four parts: 1) the statement of a conviction held, 2) a request to account for such a belief, 3) the student's own point of view, and 4) his/her supporting statements for his/her opinion (cf. Stage 2 above). This structure is expected to be represented in the content of a candidate's response and will be evaluated relying on the criterion of communication in the oral proficiency marking scale. Once the monologue has been completed or the time allotted for the monologue has elapsed (in which case the interviewer has to stop the speaker), the interviewer moves on to the questions. Each candidate is asked the same number of questions. Once again it should be pointed out that for the reasons discussed above, the interviewers should not alter the questions, nor should they leave any of the questions out, unless the candidate has already responded to the question in his/ her monologue (as stipulated in the script). The questions are constructed so that they would first tackle issues that are relevant to the candidate personally and move on to relate to events and problems related to the local community, to Estonia, and finally to global concerns, attempting as wide a coverage of the topics in the curriculum as possible.

Stage three is a role-play and should proceed according to the prototype script below.

STAGE 3: Task 2 (4–5 min.)
Interviewer: ***Here is a card with a task on it. Please read it to yourself. You have 1 minute to think about it. I'll tell you when the time is up.***

Note-taking is not allowed at this stage.
When the time is up, say ***'Could you start the role-play now.'***

Use the information in the script to answer candidate's questions.
Do not give more information than the candidate asks.
Keep your answers short and natural to oral communication.

Student's cue card
You are a journalist of a British newspaper, which is considering an article about the Pärnu Film Festival. Your interviewer is an organiser of the festival.
Ask the interviewer about
1. aim
2. time the festival started
3. organisers
4. prizes awarded
5. winner of 2006
6. time of this year's festival

At the end of the talk, say whether you think you have got enough information to write an article about the festival.

Interviewer's cue card
1. The aim of the International Documentary and Anthropology Film Festival is to learn about the culture of different ethnic groups.
2. The first festival took place 21 years ago.
3. The chief of the festival is Mark Soosaar who is assisted by many people from the Pärnu Museum of New Art.
4. The Grand Prize for the best film of the festival is a hand-woven West-Estonian blanket.
5. In 2006 the Grand Prize was awarded to Arunas Matelis from Lithuania for his film "Before Flying Back to Earth".
6. This summer the festival will take place on 1–8 July.

If the candidate does not finish the role play as required (does not give a decision at the end), ask ***'Is that all you wanted to say?'***
When the candidate has finished the role play, finish the interview by
'Thank you. This is the end of the interview.'

Switch off the cassette recorder.
Before the candidate leaves the room
• tell the candidate when the scores will be announced
• ask the candidate to sign the attendance form
• collect the candidate's notes

Stage three was transferred to the new framework intact from the previous framework, and thus the current researcher's contribution to that particular section was confined to formulating the script for the stage using the content that had already been established during previous examination development. In stage three, the interviewer has a dual role: he/she is expected to conduct the interview according to the interviewer script and not deviate from that. On the other hand, he/she will be a participant in a role-play responding

to the candidate's queries. Although cue cards are provided for the interviewer, his/her answers cannot be completely predicted as they depend on the student's wording of the questions. The interviewer as a role-play participant, consequently, would have to vary his/her answers and provide only such a response as is warranted by the question.

It has to be admitted that the role-play in its current format does not conform to the definition of role-plays as we find it in the testing literature – 'a task in a test of speaking performance in which a test taker adopts a specified role in an interaction with one or more additional speakers … and generally simulate authentic situations relevant to the communicative demands the test taker will face in real life' (Davies et al 1999:171–72). All the candidate has to do here is to paraphrase the task that has been given to him/her in a sentence and after that ask a given number of either direct or indirect questions. This kind of prompt may be helpful for a weaker student who may need all the help he/she can get to formulate a question, but might be rather restrictive for a more inventive candidate who may want to ask other questions than those specified by the prompt. It would be interesting to know what the raters' response to those candidates have been who did engage in the role-play but asked questions and made comments (linguistically and topically appropriate) that were not specified by the prompt.

3. 2. 2. A New Marking Scale for Speaking

In accordance with the adjustments made in the framework for testing speaking proficiency within the framework of the national examination in the English language, and stemming from the problems the earlier rating scale seemed to pose, discussed in the current chapter above, the author of the current research proposed a draft for a new marking scale. The author relied on the intuitive method drawing on the previous marking scales developed for assessing speaking at the national examinations in Estonia (NE 1997: 69, NE 1998: 87, NE 2000:69) as well as samples proposed by the CEFR (2001) and other international marking scales discussed in testing literature by Bachman (1990:326–28), Cohen (1994:328–32), Hughes (2003:104–5), McNamara (2000:35–46) and many others.

The scale was proposed as an analytic marking scale comprising five criteria, in which three criteria (communication, vocabulary, grammar) would be independently evaluated and the final facet would encompass two features of the speaker's production – fluency and pronunciation. The two latter were combined into one criterion in order to control the cognitive overload that keeping more than five criteria simultaneously in mind would cause to the raters (CEFR 2001:193). During the trials, there also seemed to be a positive correlation between the two features: candidates demonstrating a high level of fluency seemed also to have relatively fewer pronunciation issues. Moreover, the raters were struggling to keep the two criteria apart and thus a joint criterion was introduced in the scale. Further research is necessary as to verify if such a decision was justified, however. The scale was proposed as a full scale, i.e. a scale where all the criteria were equally weighted to avoid a negative washback effect.

The proposed marking scale in its tried and edited version can be seen below.

	Communication	Vocabulary	Grammar	Fluency&Pronunciation
5	Independent speaker. Responds to all aspects of the prompt. Interacts naturally with appropriate openings, responses, fillers and amplifications. Logical and clear. Able to paraphrase successfully.	Wide vocabulary, precise and appropriate. Word formation virtually error-free. Appropriate register.	Uses a variety of complex grammatical structures. Only very occasional mistakes.	Very fluent. Speaks fluently with appropriate pronunciation / intonation and only natural pauses. Can express him/herself confidently, clearly and politely. Often shows remarkable ease of expression.
4	Good speaker. Responds to most aspects of the prompt. Interacts mostly naturally, but may not always be logical. Responds adequately when prompted. Usually able to paraphrase successfully.	Appropriate vocabulary with occasional errors in word-formation and register. Only occasional misuse of words.	Mostly grammatical. Simple structures error-free. Complex structures are frequently attempted but these may contain errors.	Fluent speaker. Can maintain a fairly even tempo. There are occasional noticeable pauses when the speaker is looking for words. Pronunciation and intonation mostly correct.
3	Hesitant speaker. Attempts to respond to most aspects of the prompt but relies noticeably on the input with limited personal contribution. Interaction is attempted but this seems mechanical. Frequent problems with logicality.	Simple vocabulary fairly well controlled but more complicated words and expression not attempted or misused. Frequent register problems.	Relies on simple sentences only, which occasionally contain errors. Complex structure, if they are attempted, nearly always contain an error.	Hesitant speaker. Can make his ideas clear to the listener, but is not able to maintain an even tempo. Frequent self-correction hesitation and pronunciation problems lead to some misunderstanding.
2	Laconic speaker. Attempts interaction but with frequent failure. Mentions prompts, without development or ignores them. Disorganised, illogical answer. Requires assistance with little effect.	Relies mostly on quite basic vocabulary that still contains errors. Unaware of register. Words often misused. Inappropriate register all through.	Frequent grammatical errors in simple formulaic sentences. Complex structures not attempted.	Laconic speaker. Speaks with frequent illogical pauses. Unable to keep going/maintain the flow. Serious problems with pronunciation and intonation but for the most part can still be understood in spite of them.
1	Very laconic and hesitant. Unable to interact beyond mentioning the task and a rare question or monosyllabic answer.	Very limited vocabulary. Isolated words or collocations. Unaware of register.	Finds it hard to form sentences. Most utterances contain an error.	Very laconic. Pronunciation difficulties make the speech mostly incomprehensible. No traceable stress pattern.
0	Does not attempt the task. Misinterprets the task completely.	The answer is too short to allow evaluation. The vocabulary is inappropriate all through.	The answer is too short to allow evaluation. All utterances ungrammatical.	A non-speaker. The answers are too short (monosyllabic) to allow evaluation.

The above rating scale is an analytic marking scale, where the raters have to evaluate the speaker's perform distinguishing between six levels within four criteria. Even though the final score is reported as a single figure calculated as a sum of the values afforded for each criterion (thus concealing the separate ratings given for the different aspects of the performance), leading to a speculation that perhaps a holistic score, where the candidate's performance is rated as a single impression of the impact the speaker makes (McNamara, 2000:43) would have been more appropriate, an analytic scale was designed for several reasons. The first reason is the tradition that has been established, marking of the candidates' speaking ability. The markers expect and are used to working with an analytic scale. A second, more valid reason is the character of any analytic marking scale itself. Analytic marking scales, in the testing literature often also referred to as multiple trait scoring (cf. Fulcher and Davidson 2007:97), have certain advantages over holistic scoring: 'raters are required to focus on each of the nominated aspects of performance individually, thus ensuring that they are all addressing the same features of performance; it allows more exact reporting of literacy or oracy development, especially where skills may be developing at different rates (reflected in a marking profile); it leads to greater reliability as each candidate is awarded a number of scores' (Davies et al 1999:7). Research also points at particular problems that test developers and raters need to be aware of. For example, Cohen (1994) warns that 'there is no assurance that analytic scales will be used according to the given criteria; rating on one scale may influence rating on another' (1994:317). This is known as the halo effect in assessment. Cohen also asserts that subjectively marked skills (writing, but also speaking) are more than just a sum of the chosen components and individual scales may call for qualitative judgements that are difficult to make (ibid). At the same time, Weir (1990), Cohen (1994) and Luoma (2003) point out that analytic rating scales make the rater training easier. Thus the assets of having an analytic marking scale seem to outweigh the disadvantages. The third reason is an attempt in the future to align the Estonian national examination in the English language to the CEFR scale which relies on the analytic marking in many of the proposed scales for evaluation (cf. CEFR 2001).

The current marking scale is an attempt to differentiate between different levels of speaking performance within the range prescribed by the overall evaluation system used during the national examination, i.e., each skill is evaluated within the range of 20 points, resulting in the maximum of 100 points for the whole examination. As discussed above, four criteria were chosen, to control the cognitive load for the assessor, and that allowed distinction of performances on six levels. Initially, fluency and pronunciation were intended as separate criteria, but apart from resulting in rater difficulty to distinguish between the two criteria discussed above, it also would have allowed rating students only on a 5-point scale within each criterion, which did not seem sufficient, given the spread of the proficiency levels the candidates seemed to display.

The criteria in the marking scale show that the speaking construct is defined through two approaches, the linguistic approach, which focuses on language form

and evaluates task performance in terms of vocabulary, grammar and pronunciation and fluency; and the communicative approach, which concentrates on how well the candidate can use particular strategies to perform the communicative activities required (Luoma 2003:185). An attempt was made to ensure that each criterion be independent of the other criteria evaluated within the scale, to reduce the halo effect to the minimum. Consequently, the criterion of communication focuses on the content of the performance, how the speaker handles the task, the amount of the candidate's contribution, its clarity and naturalness. Vocabulary evaluates the complexity, accuracy and appropriacy of lexis, the candidate's awareness of register. Grammar assesses the complexity and accuracy of the grammatical structures the candidate resorts to while speaking. Fluency and pronunciation consider how proficient the candidate according to the two criteria is and to what extent the proficiency level here hinders or promotes the overall positive performance. The researcher tried to 'summarise descriptors into short statements to make them easy to use' (Luoma 2003:60) and to enhance the effect, key descriptors in the scale were underlined, so that the assessors would have quick points of reference to how levels differ from one another, e.g. for example Communication level 5 and Communication level 4 (independent speaker vs. good speaker). Rater training was considered essential to familiarise the raters with the level descriptors and to build up their expertise in applying the scale.

3. 2. 3. Training of Interviewers and Raters

Training of the examiners is of vital importance in order to achieve test reliability. Rater reliability can be discussed in terms of inter-rater reliability ('the level of consensus between two or more independent raters in their judgements of candidates' performance' (Davies 1999:88)) and intra-rater reliability ('the extent to which a particular rater is consistent in using a proficiency scale' (ibid: 91)). In the rating process it is important to achieve both – consistency between raters and consistency within one's own rating process from one student performance to the next. Numerous studies point to the fact that through training it is possible to achieve inter-rater reliability of 0.7–0.9 (Hamp-Lyons in Kroll 1991:79; Homburg 1984:88), which is statistically significant and suggests that readers may be guided by similar criteria in their decisions. Yet the reliability score leaves a disagreement level of 20 to 50 per cent. Vaughan's (1991) study demonstrates cases of disagreement between raters in up to 3 points on a 6-point scale (Vaughan 1991:115), which in most cases is a difference between pass and fail. The readers in the study were all trained and experienced, allegedly guided by the same (stated) criteria for marking.

Alderson et al (1995) divides the training of raters into the following stages: designing the rating scale, setting the standards (choosing a number of sample recordings to illustrate each level of performance according to the rating scale) and the standardisation meeting. The aim of the standardisation meeting (sometimes also referred to as

moderation meeting, e.g. McNamara 2000) is manifold: familiarisation of the raters with the marking scale, so that they are familiar with the criteria and the different levels within each criterion. This is followed by marking the so-called consensus speaking samples – the speaking samples that can be related to a particular score on the rating scale without much disagreement between raters. The final stage Alderson et al (1995) suggest is marking problematic speaking samples – speaking samples that do not seem to fit any of the criterion levels exactly. The aim of the standardisation meeting is to reach agreement about what each performance is worth and how to apply the rating scale should a problem arise (110–115).

Apart from training rating per se, Alderson et al (1995), Bachman (1996), Hughes (2004), Fulcher and Davidson (2007) discuss the standardisation of the procedures and the need to train test administrators. 'The examiners should be given instructions about where to sit in relation to the candidates, what kind of questions to ask in order to bring out the best in the candidate, how to manage the many papers that they will be holding (not only their own instructions, rating scale and scoring sheet, but also all the material that the candidate will need to refer to), how to enter their marks discreetly, how to welcome the candidate and bring the test to a close, and so on. (Alderson et al 1995:115).

In order to train the examiners and assessors for the speaking section of the national examination in the English language in Estonia, the existing procedural guidelines both for the examiners and assessors were thoroughly revised and updated (cf. Appendix 2) which prescribe the examiner and the assessor behaviour in terms of preparation for the interviews on the examination day, time and document management, the amount of examiner support available to the candidate, affording and recording of scores, etc. In addition to that, all teachers scheduled to act as either assessors or interviewers or both during the national examination in the spring of 2008 had to go through a qualification training with the National Examination and Qualification Centre. The training course followed the outline proposed by Alderson et al to which training in examination procedures was added. The interviewer training is a 4-hour programme, consists in familiarising the interviewers first with the concept of scripts (their definition and function), introducing the framework of the speaking test along with the time-line, i.e. walking the interviewer through the test situation. Additional information is given about the organisation of the room, the position of the interviewer and the assessor, the necessary documentation and how to handle it, along with a code of behaviour for the interviewer at various stages of the interview. The interviewers practice conducting separate parts of the interview (introduction, task 1 and task 2), watch a training video of the interview which they comment on relying on a mark-and-comment sheet and finally carry on a full-scale interview with a colleague acting as a candidate. The training video, the marking scale and the interviewer guidelines are available for the interviewers in training on the Internet to make frequent recourse to. Practical (financial) constraints control the production of additional training materials and employment of volunteer candidates. As most of the people signing up for the interviewer training also act as assessors on a different day of the national examination,

a separate part of the training is connected with the discussion of the marking scale, listening to sample recordings for standardisation and then marking some of the recordings independently.

3. 3. INTERVIEWER / RATER QUESTIONNAIRE

The new interview format was expected to cause a number of procedural problems. The majority of the examiners in Estonia were mostly unfamiliar with the concept of scripts and a certain amount of reluctance was expected with regard to their implementation during the interview. This involved first and foremost adhering to the wording of the scripts, as procedural comments repeated unchanged from one interview to the next may have seemed artificial and boring to the interviewers. This may have resulted in interviewers paraphrasing, amplifying or summarising the information given in the scripts, which in its turn would result in the candidates getting a varying amount of input during the interview. A concern related to this involved cases where the student requested additional information, either a clarification of the statement given as prompt for the monologue or a definition of a word or an expression they claimed not to know.

The second concern involved keeping time. Earlier spot checks of the recording of the interviews had demonstrated a variation in the interview time of up to ten minutes at times, which made it impossible to compare the interview results and, in effect, deprived them of generalisability and the interview itself of reliability and validity. As discussed above, time also initially seemed a concern when setting limits to the candidate monologue in task 1 and the candidates' preparation time for the monologue. A 2-minute time limit for the monologue was set, relying on the examples provided by international proficiency exams most frequently chosen by Estonian students in pursuit of continuing their studies abroad after gymnasium/upper-secondary school. Preparation time for the monologue was established at 3 minutes to allow students sufficient time to plan their answer on the one hand, and to avoid resorting to writing out their complete response and then reading it out during the oral examination, on the other. It was now necessary to establish to what extent the examiners adhered to the time-limits set.

A third concern involved interviewer language while conducting the interview. Information was sought as to how efficient the interviewers were in making smooth transitions from one interview stage to the next: stopping the candidates when necessary, introducing the next stage, etc. Seamless transitions contribute to reducing stress during the interview, which in its turn may help the candidate to display his/her speaking proficiency at its best.

Research also tried to establish the interviewers' perception of how the students seemed to respond to the new national examination speaking section format, its content as well as its procedure, what the reaction to the monologue topics was if they were age and gender appropriate.

In addition to the issues above, feedback was necessary with regard to the marking scale, its content, appropriacy and ease of use.

Finally, there were concerns about certain other practicalities regarding the oral interview, the availability and implementation of recording, the general atmosphere of the environment of the interview and to what extent the latter could be controlled in each particular case by the examination administrators.

The new interview format was first implemented during the national examination in the English language of 2008, with optional recording of the interview proceeding from the Estonian Ministry of Science and Education regulation. All in all, 9293 students took part in the oral interview and 624 opted for recording.

Answers to the above-mentioned concerns could be found resorting to two kinds of procedures: by conducting a questionnaire among the teachers who acted as either interviewers or assessors or both during the 2008 national examination and by analysing the recordings of the interviews with the candidates at the national examination. The final section of this chapter will discuss the interviewer/assessor questionnaire. The analysis of the recordings will be taken up in the subsequent chapter.

In order to find out the interviewers' reactions to the new procedure, a questionnaire study was carried out among them. The questionnaire originated from the need to investigate interviewer response to the new oral proficiency interview framework. The questionnaire was developed relying on the interviewer training programme, by choosing the most salient features of both the training programme and the OPI. Because each element in the interviewer script was carefully considered during its design, and very precise instructions were given to the interviewers during training concerning the use of language, time-keeping, note-taking etc. during the interview, the questionnaire also aimed to be as specific as possible about the different interview aspects, contrary to earlier questionnaires seeking interviewer/rater feedback about the English language national examination exclusively in very broad terms. Thus a number of statements were drawn up describing as many aspects of the interview as possible. The number of those statements was finally reduced to 40 for practical considerations, for fear of the interviewers losing their motivation if the number of them was too large. The questions were designed to elicit data about the interviewers' perception about their preparedness level to conduct the interviews: the amount and quality of training that they received (statements 1 to 6), usefulness of a script during the interview (statements 7 to 11), the effort necessary to keep and manage time (statements12 to 16), their perception of various aspects of student behaviour during the interview (statements 17 to 25), the quality of task 1 and task 2 (statements 26 to 35), their attitude to the marking scale (statements 36 to 37), their anxiety level and practices concerning recording the interview (statements 38 to 39) and the examination room set-up (statement 40). The statements were reviewed and revised by the NEQC examination development staff, and alterations were made to some of them. The questionnaire was initially designed to contain just the statements with the Likert scale, to be completed by the interviewers. The respondents had to record

their opinion concerning the statements in the questionnaire on a 5 point scale ranging from 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. The dangers that are sometimes linked with the scale – central tendency bias, acquiescence bias – were anticipated, but it was hoped that by questioning the interviewers anonymously, these problems would be somewhat alleviated. To further combat this, and to cater for any other unforeseen interviewer observations, a section was added to the questionnaire where the respondents were asked to comment on the training materials, exam procedure, marking scale, and any other aspect of the exam that they felt needed commenting. For the full questionnaire form see Appendix 3.

As additional information, the respondents had to specify the amount of experience they had had teaching in the upper-secondary/gymnasium level and for the purposes of this research were initially divided into six groups, those with one to two years of experience, those who had between 3 and 5 of experience, those who had taught between 6 and 10 years, those with 11 to 15 experience, those with 16 to 20 years of teaching experience and those who had more than 20 years of experience with the above-mentioned school level. The distribution of the working experience of the questionnaire participants was the following: 1–2 years – 12 participants, 3–5 years – 9 participants, 6–10 years – 9 participants, 11–15 years – 12 participants, 16–20 years – 18 participants, more than 20 years – 21 participants. The number of people, however, in each subgroup proved to be too small to allow any systematic generalisations with regard to a relationship between teaching experience and a particular behavioural pattern. Thus, for the purposes of the current investigation, this information is of limited value, and serves as a possible starting point for further more focused research on the impact of teaching experience on interviewer behaviour.

The criterion for including the people in the study was that the participant had passed the pre-examination training and that he/she had actually interviewed students during the oral section of the 2008 English national examination. The respondents were teachers who took part in a series of teacher education sessions carried out by the researcher in the northern (Tallinn) southern (Tartu) and western (Pärnu) part of Estonia. The participants were all volunteers. The questionnaire form was handed to the teachers by the researcher and the respondents completed it on site. All in all, 82 questionnaire forms were issued of which all were returned and 81 proved to be usable for the purposes of this study.

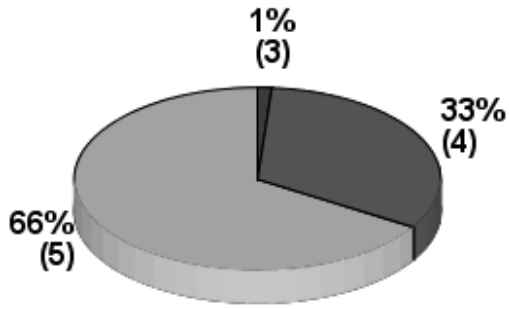
The questionnaire responses were analysed with statistical methods. In addition to descriptive statistical analysis, cluster analysis was completed with the aim of establishing whether groups could be found who would display similarities in their disposition, and Spearman rank correlation was resorted to in order to possibly relate the disposition of the respondents to their teaching experience. Statistical analysis was performed data processing program SPSS v 16.

First, the results obtained will be presented relying on descriptive statistical analysis.

3. 3. 1. Preparedness Level for the Interview and the Quality of Training

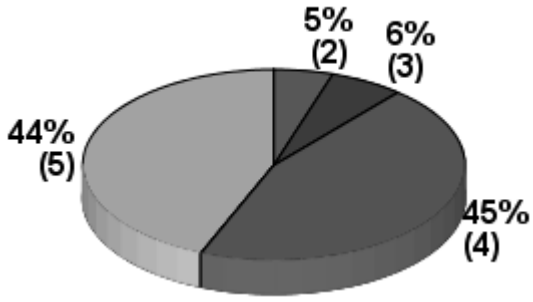
Statements 1 to 6 in the questionnaire pertained to the interviewer's perception of their preparedness level for the interview and their satisfaction with the training.

Figure 1. Q1. I was clear about the exam procedure before the exam started. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



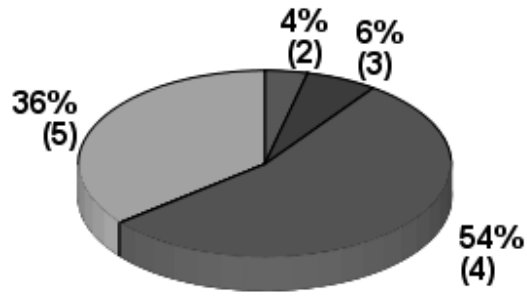
The respondents report being either absolutely (66%) or mostly (33%) clear about the examination procedure, with only 1 per cent finding it hard to comment on the statement. This may reflect the actual state of affairs or may just demonstrate the respondents' awareness of the need to be clear before starting the interviews. The validity of the claim would have to be corroborated by studying the recordings of the interviews to see to what extent the procedure was observed. On the other hand, the fact that the teachers report confidence about their command of the procedure may be an indication of the fact that they will attempt to follow the procedural requirements and thus contribute to producing reliable test results.

Figure 2. Q 2. Pre-exam training was sufficient. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



Before interviewing the students, all examiners had to participate in a 4-hour instructor-led training session, which consisted of a theoretical introduction and explanation of the new examination pattern and a hands-on practical session where the examiners had an opportunity to try the procedure out, ask questions and offer comments. This was supplemented by watching a training video, which featured the new procedure, and commenting on the procedure relying on checklist of pertinent features. The final part of the training consisted in evaluating audio-recorded student performances. All the training materials were also posted on the National Examination and Qualification Centre (NEQC) website, with instructions to teachers to practice more independently. Questionnaire participants stated that the training was either absolutely (44%) or mostly (45%) sufficient. 6 per cent were unable to express an opinion and 5 per cent of the respondents found the pre-training mostly insufficient. It is important to remember here that this is the teachers' opinion of what is true. Further training needs would have to be established based on how the teachers actually managed the interview.

Figure 3. Q. 3. The examiner materials were helpful. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

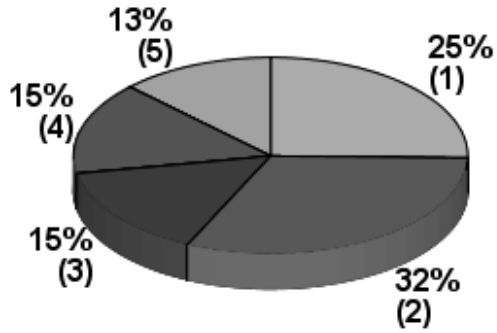


Examiner materials differed from the materials the interviewers had previously had in that this time, they required close attention and allowed little deviation. On the one hand, this meant that the interviewers had to spend extra time familiarising themselves with the material and make an effort to adhere to them. On the other, the scripts freed the interviewers from the worry of finding the appropriate language to guide the candidate through the interview. The respondents seemed to predominantly value having scripts as 36% claimed them to be absolutely helpful and 54% found them mostly helpful. 3% of the respondents could not decide and 4% found them mostly not helpful. It would be useful to design further research with the latter to find out why this was the case.

Figure 4. Q.4. I would need more training in the exam procedures. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

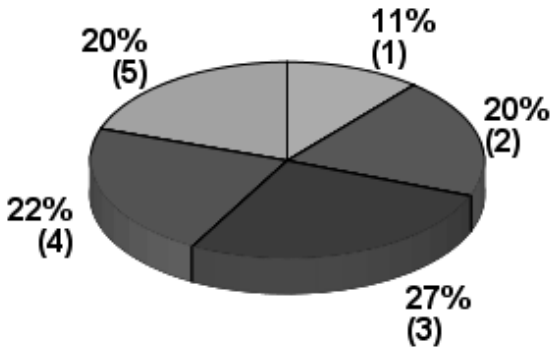
The answer to this question demonstrates a greater degree of ambiguity among the respondents. Although question 1 seems to indicate that the vast majority of

the respondents were clear or mostly clear about the exam procedures (99%), their perception of needing further training varied quite notably.



Thirteen per cent of the respondents were absolutely convinced they needed further training and a further 15 per cent thought they probably needed it. On the other hand, 25 per cent did not see any need for further training and 32 per cent noted this was mostly the case. 15 per cent of the teachers were unable to respond. The answer may reflect the dichotomy between their knowledge about the procedure and their skill to implement it, they know what the procedure should be like but they need more practice in the execution of it.

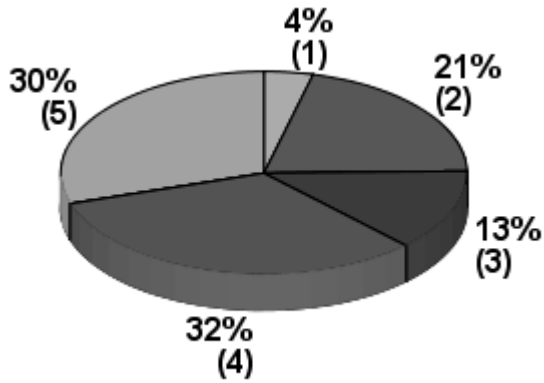
Figure 5. Q. 5. I am willing to take part in an additional training course. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



This question shifted the focus from the instruction and information provided by the examination development team to the teachers' own initiative, their willingness to take on additional responsibilities. 20 per cent of the respondents were prepared to take part in additional training courses and 22 per cent claimed to be almost certain to be willing to do so. 27 per cent could not decide, whereas 11 per cent were absolutely not willing to participate in further training and 20 per cent report that they

are almost certainly not ready to do so. So about a half of the teachers questioned perceived a need for further professional development with regard to the examination procedures.

Figure 6. Q. 6. I check the examination centre website frequently for new materials. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

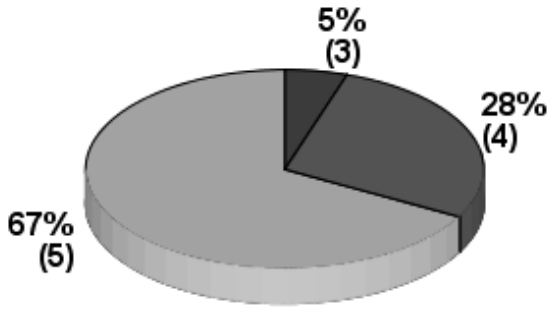


This question was set to get further information about the teachers' own initiative while seeking national exam related public information and finding ways of self-development. Questionnaire participants demonstrated quite a variety of behaviour. Thirty per cent of the respondents claimed to check the examination centre website and 32% said this was mostly true. Four per cent said they never did it and 21 per cent said they mostly never did it. Thirteen per cent found it hard to respond to the question. This may suggest that one third of the teachers do not make use of the training materials and the public information on the NEQC web-page as a source for professional development that is available to them as teachers, examiners and assessors.

3. 3. 2. Usefulness of a Script during the Interview

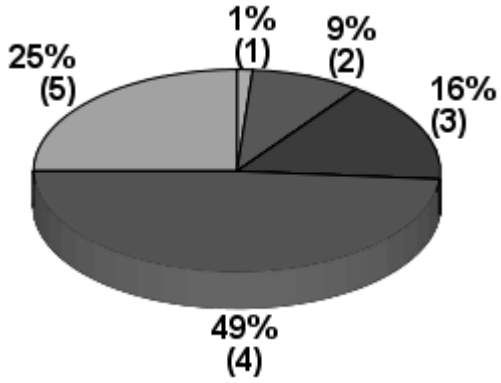
Statements 7 to 11 tackled the issue of using a script during the interview. As discussed above, it was for the first time that the interviewers had to rely on a word-for-word script during oral interviews. Discovering the interviewers' attitudes towards them was considered important from the point of view of test reliability. If the teachers recognised their value as a means of enhancing test reliability, they would be more likely not to deviate from them. If, on the other hand, the teachers did not recognise their function, more training of interviewers would be necessary. Also, if the interviewers did not perceive the need for interviewer standardisation, it was highly likely that their interviewing practice would reflect that and give rise to student responses of different lengths and of varying examining conditions for the students.

Figure 7. Q. 7. Having a script for the interview was helpful. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



As can be seen from the chart above, only five per cent of the teachers could not decide if the script was helpful or not. The overwhelming majority of the teachers considered having a script either mostly (28%) or absolutely (67 %) useful. Thus the speculation that teachers needed help in the examination procedure seemed to have been corroborated by the responses to this question.

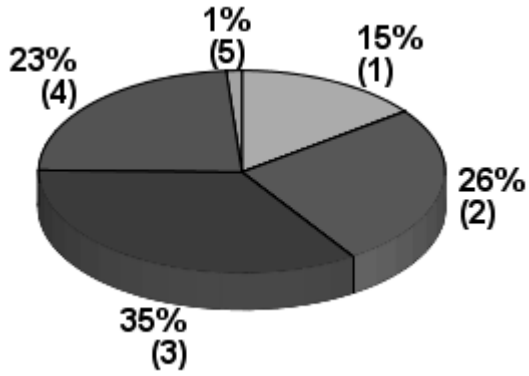
Figure 8. Q. 8. It was easy to keep to the wording. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



One of the interviewing skills that was practiced during training was adhering to the script wording with all the candidates. Altering the wording was perceived by the test development team as a means of changing the tone and modality of the instructions/comments/questions, which would consequently also lead to the students potentially being in unequal testing conditions. This question was set to get further information about the teachers' own initiative while seeking national exam related public information and finding ways of self- development. The figure above demonstrates that about three quarters of the respondents found it (25% absolutely, 49%

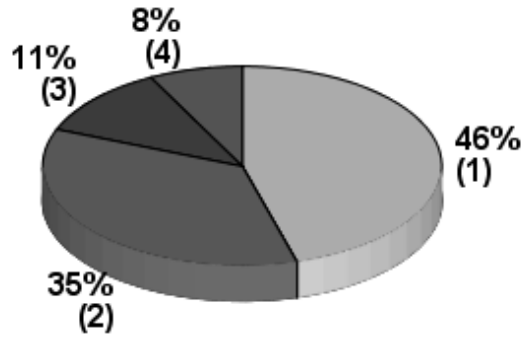
mostly) easy to follow the wording. That does, however, leave slightly more than a quarter of those teachers who had a problem following the wording (1% absolutely not easy, 9% mostly not easy to follow the wording and 16% unable to decide). Based on this question, it would be difficult to say anything about the actual adherence to the script. Both, those who found it easy to keep to the wording and those who found it difficult, may have, in actual fact, either still adhered to it although it was difficult or deviated from it although adhering to it was easy. Further investigation is needed as to how the interviewers actually behaved during the interview as well as the reasons why following the wording was difficult.

Figure 9. Q. 9. The wording of the frames seemed artificial. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



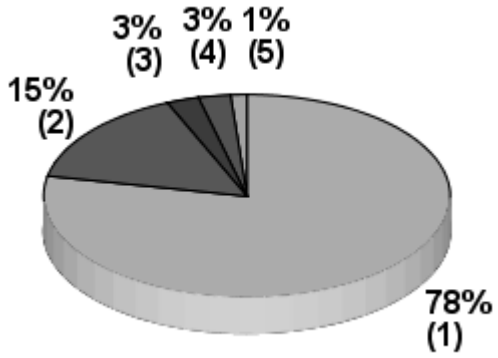
Here, the biggest proportion of respondents fall in category 3 (hard to say) – 35% – which may indicate that group's insecurity about their own language competence, and leads them to abstain from passing a judgement. It may also signal the groups unawareness of the language generally employed in that particular context. It may also be that as teachers they are seldom asked to comment on the naturalness of the language they encounter in national examination (and other) documents issued by the NEQC. The responses suggest, too, that 1 per cent find the language of the frames to be completely artificial and a further 23 per cent consider it somewhat artificial. It would be interesting to get concrete examples and comments as to what exactly seemed artificial about the language. One might speculate here that it was not so much the language as the procedure itself that appeared artificial, since none of the teachers had ever had to maintain an unaltered interviewing pattern during oral interviewing before and to some of the interviewers this might have seemed tedious and boring and consequently, artificial. It may also have something to do with the age of the respondents, but that would need further substantiation through a more focused study. Thus the new testing practice may be too much at odds with their habitual teaching and testing practices to appear natural.

Figure 10. Q. 10. I changed the wording of the script. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



However artificial the wording of the frames might have seemed, the figure above demonstrates that the vast majority of the group did not consciously either ever change the wording of the script (46%) or almost never (35%) did so. Eight per cent admit sometimes having changed the script. The most interesting group of the respondents is the 11 per cent of the teachers who found it hard to answer the questions, which may indicate that keeping to the wording of the script was not something that they considered particularly important, in which case it is more likely than not that their adherence to the script was random.

Figure 11. Q. 11. I ignored the wording of the script completely. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



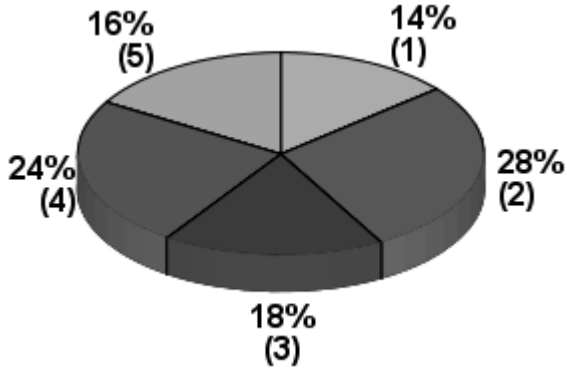
Seventy-eight per cent of the study participants claim they did not ignore the wording. That, however, leaves 22 per cent of the teachers who deviated from the script quite consciously. This in its turn means that more than one fifth of the candidate population was being interviewed under very different conditions compared to their peers and consequently may have been subjected to severer or easier interview conditions.

Further investigation is needed to see if indeed examiners adhered to the scripts and if not what sorts of changes were made as well as the reasons why the changes were introduced.

3. 3. 3. Keeping and Managing Time

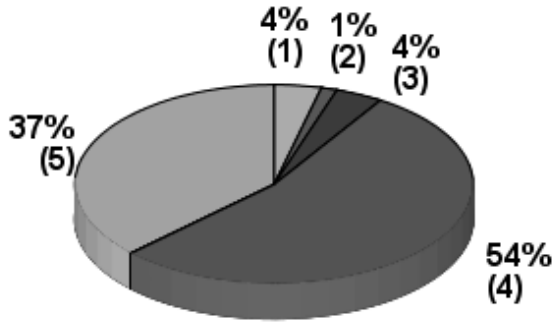
Subjecting students to time constraints during the interview was another means in the effort to achieve comparability of interview results and consequently, reliability of examination results. Statements 12 to 16 looked at time keeping and management.

Figure 12. Q. 12. Keeping time required effort. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



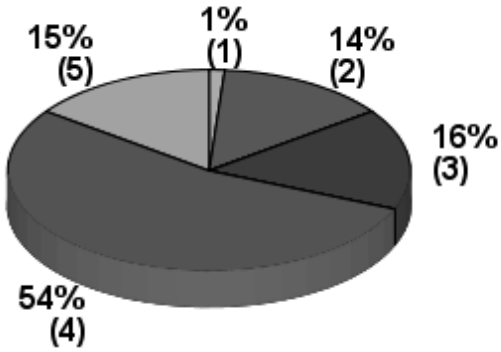
It can be seen that the teachers are quite evenly divided here with respondents falling in all the five groups. Slightly more than a third of the study participants denied having absolutely (14%) or mostly (28%) no problem with time keeping. About the same amount of respondents admitted time-keeping to be a serious (16%) or a somewhat serious (24%) problem. Eighteen per cent of the respondents could not answer the question, which may indicate that they did not pay sufficient attention to imposing uniform temporal conditions on the interviewees which again may have lead to variation in the exam results due to unequal conditions. Imposing a time constraint on oral interviews is a common practice that allows comparability. It is a feature that is present at other international proficiency examinations that Estonian students may have to take to proceed to other educational institutions. Thus it is important that both students and teachers learn to work under time constraints. Ignoring this aspect of oral interviews will not only deprive the oral examination results of reliability, it will also give the students a false impression of how their oral proficiency will subsequently be assessed should they take similar international examinations and may lead to very different oral proficiency estimations.

Figure 13. Q.13. I kept to the required timeframe. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



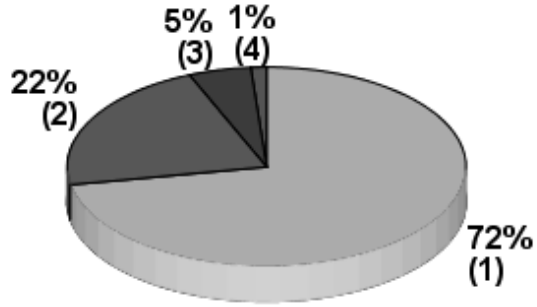
During training, temporal considerations were among some of the most discussed aspects of the training procedure. All the teachers who participated in the study, had previously been trained, and were thus aware of the requirement. When questioned, if they had kept to the required timeframe, the overwhelming majority maintained that they either absolutely (37%) or mostly (54%) did keep to the timeframe. This answer, first and foremost demonstrates the teachers' awareness of the requirement. If they actually did so would have to be verified through listening to the recordings of the interviews. Five per cent of the respondents say that they either never (4%) or mostly never (1%) followed the timeframe for some reason. A further 4 per cent could not answer the question, which again is likely to suggest that they ignored the requirement. Hence there is further corroboration to the concern that there was a proportion of students whose spoken sample was collected under circumstances that differed from those of the rest.

Figure 14. Q. 14. It was easy to stop students. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



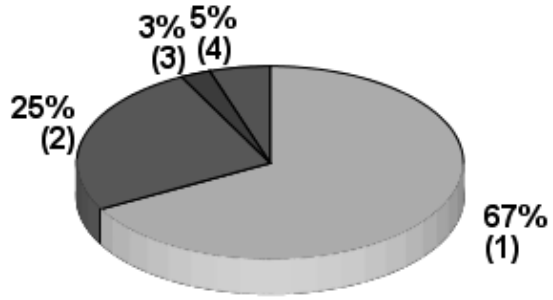
One aspect of managing time during the interview is noticing when the student monologue needs to be wound up and actually having the language to do so without sounding unduly abrupt, i.e. how to make a smooth transition from the monologue to the questions while staying within time-limits. The chart above demonstrates that 15 per cent of the teachers did not see any difficulty in stopping the student's monologue and 54 per cent asserted mostly not having problems with that. It would seem, however, that the rest of the group would need either more practice or more training to feel more confident about keeping the students within the given time since 1% of the respondents noted having serious problems with stopping the students and 14 % said they had had some problems with that. Sixteen per cent of the respondents could not answer the question, which could be an indication that they had not stopped the students monologue and let them speak as long as they wanted, or had not considered how they had made the transitions.

Figure 15. Q. 15. I forgot about the time. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



The respondents were quite clearly predominantly aware of having to keep the interview time under control with 72 per cent responding that they never forgot about the time and 22% saying that this was mostly the case. The 5 per cent who found it hard to respond may be among those who do not regard the time factor to be very important in the oral interview.

Figure 16. Q. 16. I let the students talk for as long as they wanted. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

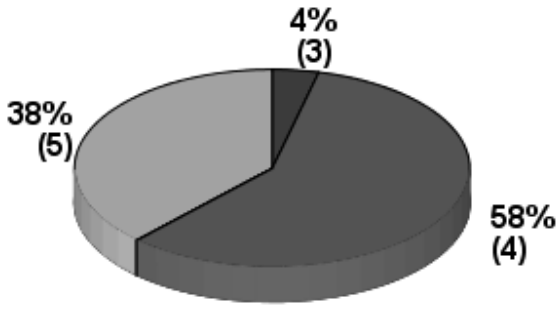


The results of the responses to this statement seem to corroborate the claims made earlier by the teachers that they mostly did try to monitor the time during the interview. The overwhelming majority assert that they either never (67%) or mostly never (25%) let the student continue as long as they wanted. Five per cent of the teachers often did so and 3 per cent were unable to answer the question. Here, it would be interesting to know what aspects of candidates' performance prompted the interviewer to extend the time envisaged, how much more time the students seemed to need to communicate his/her ideas and what proportion of the students could actually not produce a rateable sample within the two minutes.

3. 3. 4. Student Behaviour during the Interview

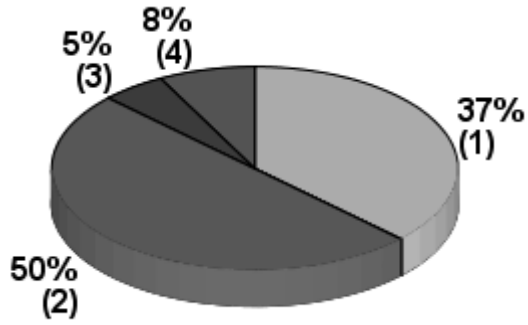
In addition to finding out how the teachers responded to the new interviewing procedure, information was requested concerning the students' behaviour during the interview as perceived by the teachers. Statements 17 to 25 studied the interviewers' perception of the strategies the students used to prepare for the monologue and to cope with the time allotted for delivering the monologue.

Figure 17. Q. 17. Students understood what they had to do. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



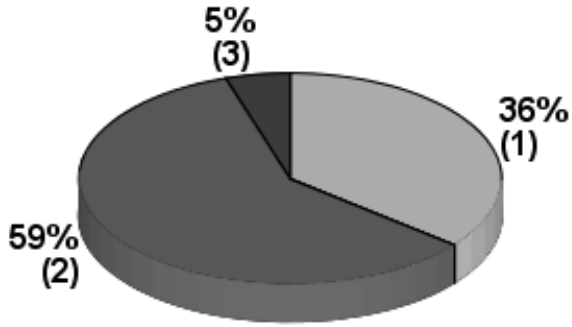
While putting together the examination procedure, every attempt was made to try and make the sequence of the procedures as logical as possible and to guide the students through the procedures with language that would be clear, concise and easy to understand. The chart above reflects the teachers' perception as to whether the students comprehended their task. Thirty- eight per cent of the respondents were absolutely certain that the students understood the task and 58 per cent stated this was mostly the case. Four per cent could not comment on the subject. It seems to suggest that the task developers' efforts to produce a clear task seemed to have been successful, yet begs the question to what extent the teachers' impression can be trusted. They made no comment as to what their positive impression was based on, and one can only speculate here that the impression was either made based on how readily the students started to respond to the task or how the students commented on the task after the interview.

Figure 18. Q. 18. Students asked you to clarify what they needed to do. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



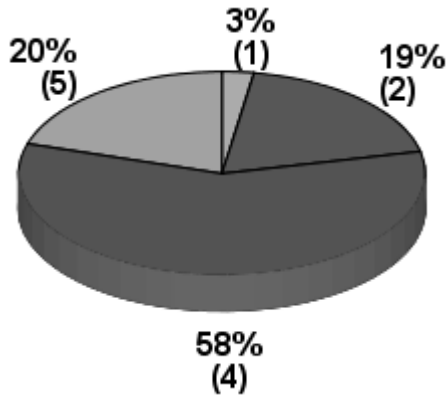
The response to this statement sheds a slightly different light on the issue of task clarity. We can still see that the majority of students seemed to teachers to either never (37%) or mostly not (50%) need additional explanation. As with the previous question, the respondents claimed that the students were overwhelmingly clear about their task, but this set of responses leads us to believe that this may not always have been the case, as 8 percent of the respondents report students needing clarification most of the time. Now the question arises what form this request took (a statement of not understanding, a request to translate, etc.) and what aspect of the prompt seemed ambiguous (the problem per se, a word, a procedural aspect). It is also intriguing why it was hard for some teachers (5%) to confirm if the students asked them to clarify anything about the task or not.

Figure 19. Q. 19. Students asked you to explain words. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



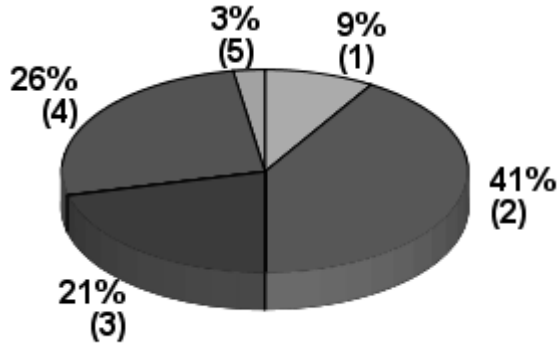
This statement probed the issue further by enquiring about the students' reaction if they encountered unfamiliar vocabulary. As can be seen from the table above, the students seemed to either never (36%) or hardly ever (59%) to request word explanation. This does not mean however that vocabulary problems did not exist. The students may not have asked for an explanation but for a word translation; so in their survey responses, teachers responded as discussed above. The 5 per cent of the respondents who noted that it was hard to say if the students did ask for an explanation or not may have struggled with the interpretation of the statement (does word translation count as explanation?) themselves and thus marked the statement as hard to comment. A question also arises as to the meaning of 'hardly ever', how frequently explanation had to be sought for the teachers to opt for that response. If, on the other hand, the difficulties in understanding, reported in statement 8, were not vocabulary-induced, further investigation is needed as to what did induce them.

Figure 20. Q. 20. Students took notes. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



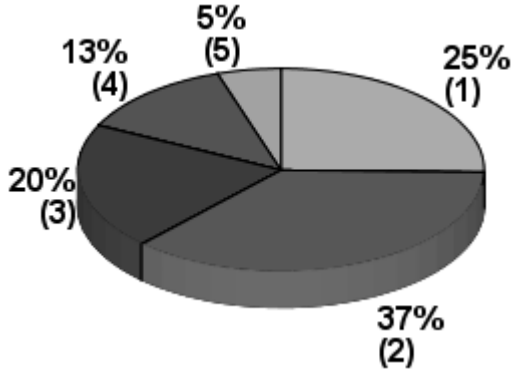
During the pre-examination training, teachers were advised to encourage the students to make notes in the 3-minute preparation time for their proposed monologue to give their presentation more structure and to make sure that they would indeed have enough to say during the 2 minutes given to them for the monologue. It was, however, not a requirement and students could forgo that tool. The teachers' responses in the graph above show that note-taking was quite widespread – 20% report this to be absolutely true and 58% saying this was mostly true – yet close to a quarter of the respondents noted that, with their students, this trend was not prevailing (19% – mostly not true and 3% – absolutely not true). As stated above, note-taking was not obligatory, but a further study could be carried out as to whether there would be a qualitative difference in the students' oral examination responses when they did take notes and when they did not do so. Based on that, more substantial recommendations could be made regarding students pre-examination training.

Figure 21. Q. 21. Students were ready before the given preparation time. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



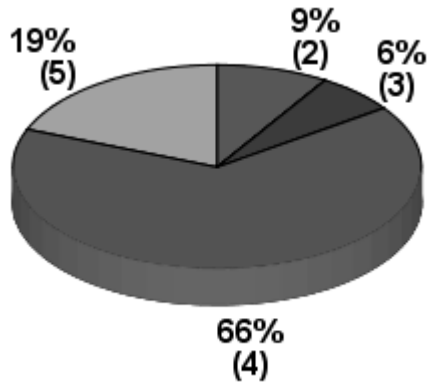
The current exam's 3-minute time limit was mostly set based on the negotiations with the teachers of English in Estonia and on the pre-testing experience. Half of the respondents (9% absolutely and 41% mostly) claimed that the students needed all the given time set aside for them for monologue preparation. Slightly less than a third confirmed that their students (3% absolutely and 26% mostly) were ready to commence with the monologue before the preparation time had expired. Quite a large proportion (21%) could not comment on the issue. It would be interesting to investigate the reasons for the latter, as at this point one is tempted to speculate that the reason for that difficulty was the interviewer's neglect to observe the time constraints set for the procedure.

Figure 22. Q. 22. Students required more time than they were given. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



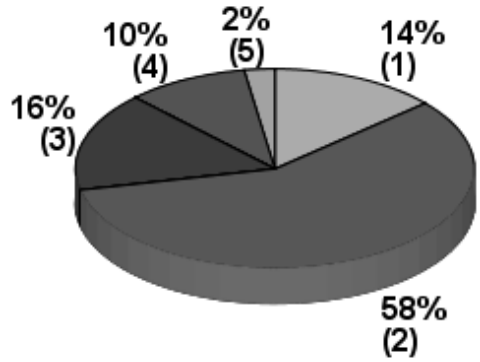
The respondents seemed to be quite divided here. The majority of the respondents denied perceiving their students to need additional preparation time (25% absolutely not, 37% mostly not). About a third of the teachers indicated that students either certainly (5%) or mostly (13%) needed more preparation time. Further research is needed as to whether the students who seemed to have needed more time actually did so and if they would have produced a better monologue given more preparation time. The amount of respondents who failed to comment on the issue coincides with those in the previous group and may denote a group who ignored the time and let the students spend as much time for the monologue preparation as they wanted. It would also be interesting to ask students if in their opinion having little time contributed to an increased stress level. If that was the case, more research would be warranted about the need to prolong the time allotted. If, on the other hand, a reliable speaking sample was obtained relying on the 3-minute preparation time, the need to extend the preparation time would be questionable.

Figure 23. Q. 23. Students used all the 2 minutes for the monologue. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



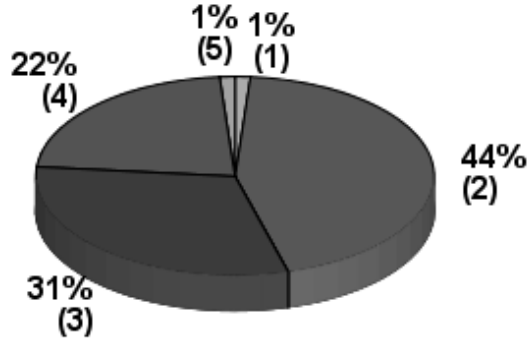
It was estimated during the task development process that, in order to complete the given monologue task adequately and get a rateable sample of student' speaking ability, between one and two minutes were needed. As we see in the chart above, in the teachers' estimation, most of the students could keep talking for that amount of time (19% absolutely true, 66% mostly true). Nine per cent of the interviewers claimed it mostly not to be the case, which in other words probably means that they finished speaking before 2 minutes were over. Six per cent of the respondents could not comment on the statement, which may mean that the teachers were so involved with the students' responses or their own forthcoming duties as interviewers that they failed to notice the time.

Figure 24. Q. 24. Students finished before the 2 minutes were over. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



This graph, too, confirms that, in the teachers' estimation, students used all the allotted time for the monologue (14% absolutely, 58% mostly). The proportion of teachers whose students finished ahead of time is relatively low (2% absolutely, 10% mostly). It would be necessary to estimate what proportion of students they represent, and if, in spite of the shorter monologue a rateable sample of those students' spoken language was still obtained.

Figure 25. Q. 25. Students wanted to talk more. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



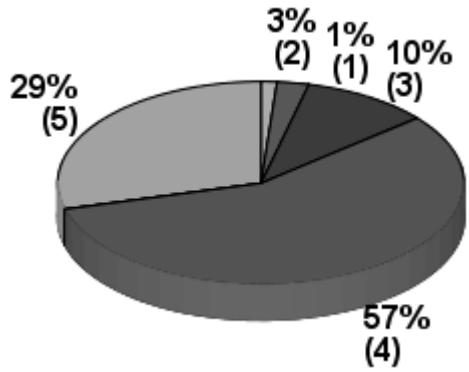
Although 2 minutes is usually sufficient to provide a representative sample of sustained speech, students may feel that they had not been given a fair chance to demonstrate their speaking ability if the time allotted for the monologue is too short. If there is an overwhelming feeling that this was not the case, monologue time could be extended. It would seem, looking at the data above, that to a very great extent this was not the case (1% absolutely, 44% mostly). Nearly a quarter of the respondents

(1% absolutely and 22% mostly) thought the students did want to talk more. Correlations would have to be established between the students' speaking score and the scores obtained in other skill areas, to show that they were not disadvantaged during the oral interviews (if the score for speaking is considerably lower that the scores afforded for other skills, analysis would be needed if this did not stem from the inability to demonstrate their proficiency due to lack of time). Students' willingness to talk longer than their allotted time may also be an indication of the appropriate task difficulty, which allowed them to give a sample of their speaking proficiency without undue effort. Thirty- one per cent of the teachers were not able to comment on the topic probably feeling not in a position to speculate on the students' point of view. To get a more trustworthy opinion on the issue, students need to be asked the above question. This could be one of the directions for further research.

3. 3. 5. The Quality of Tasks 1 and 2

Statements 26 to 35 were designed to find out the interviewers' assessment of the quality of tasks 1 and 2 and the interviewers' own account of their behaviour while administering those tasks. The aim of the tasks was not to evaluate the students' content knowledge, i.e. what and how much they knew about the history, geography, culture and politics of English-speaking countries but to find out to what extent students were in control of the linguistic, socio-linguistic and pragmatic material specified in the national curriculum and Year 12 Handbook (specifications), and to do so through easily identifiable problem situations and role-plays that would have relevance to their own lives.

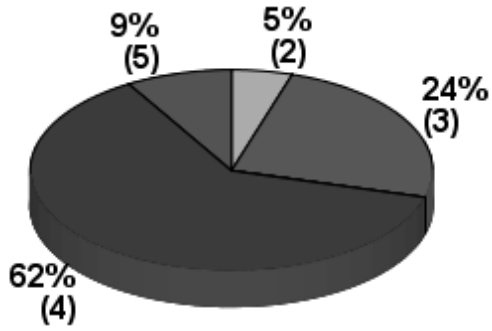
Figure 26. Q. 26. Monologue topics were easily understandable. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



Monologue topics were perceived to cause no particular difficulty (29% absolutely, 57% mostly). Just a small proportion of the teachers reported topic difficulty (1% absolutely, 3% mostly). Unfortunately, no comment has been offered as to the nature of the difficulty.

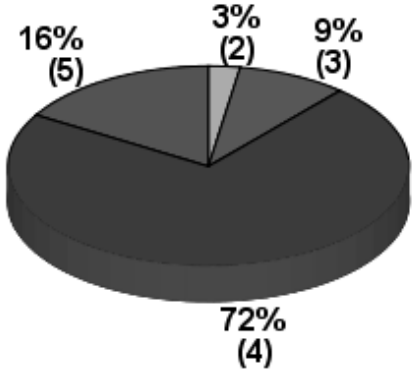
Ten per cent of the respondents reported ambiguity on the issue but for future improvement of the monologue topics, additional interviews with the above respondents (group 1,2 and 3) could be conducted to establish the precise nature of the problem.

Figure 27. Q. 27. Students found it easy to express their opinion on the topics. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



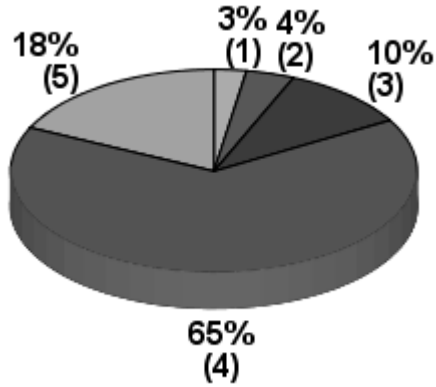
The interviewers' impression of the students' performance from the above point of view was that in the overwhelming majority of cases this was either absolutely (9%) or mostly (62%) true. This hopefully testifies that the problems for the candidates to discuss during the monologue were such that readily allowed them to draw on their life experience and their general reading and, consequently, to display their speaking ability. The remaining respondents fell into two groups: 5 per cent of all respondents that their students in their opinion mostly did not find it easy to express their opinions and a fairly large group of respondents (24%) who could not comment on the subject. Further research could establish what the nature of the perceived difficulty was.

Figure 28. Q. 28. Monologue topics were age appropriate. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



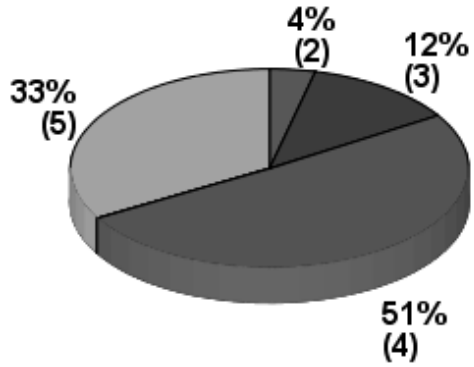
While selecting monologue problems, conscious efforts were made to avoid problems that this particular age group (students who are mostly between 19 and 20 years of age) could find it hard or inappropriate to tackle. Examples of discarded problems comprised those that required opinions about university life (which the students were not yet familiar with), retired people's problems (which they were too young to fully discuss), but also problems related to religious and racial topics. The teachers' opinion as to the age appropriateness of the topics seems to be predominantly favourable in that 16 per cent state that they were absolutely age appropriate and 72 per cent say that they were mostly age appropriate. Only 3 per cent of the respondents report that the problems were mostly inappropriate, without unfortunately providing any clarifying comment. Nine per cent fail to comment on the issue for reasons that they have not disclosed but that might be connected to being clear about the definition and understanding of the term.

Figure 29. Q. 29. Monologue topics were gender appropriate. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



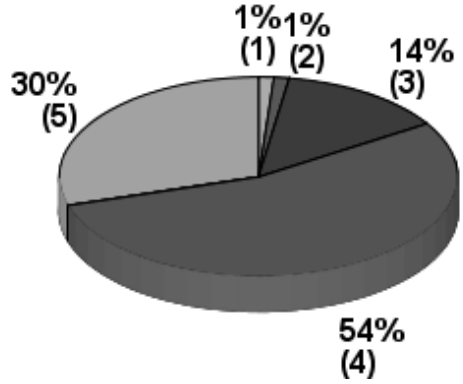
Research has shown that particular topics might be restricted to being easily accessible to only either male or female students. Attempts were made during examination development to insure that the topics would not be restrictive. The figure above indicates that the teachers predominantly considered the monologue topics gender appropriate (18% absolutely, 65% mostly). Three per cent of the respondents found the topics absolutely inappropriate for their students and 4 per cent considered them mostly inappropriate. No clarifying comments were offered as to what makes them such or which topics seemed to be inappropriate. Te per cent of the teachers could not decide if they were appropriate or not which may demonstrate that they are not aware of the concept of gender appropriateness or have not consciously thought about it.

Figure 30. Q. 30. The follow-up questions were helpful. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



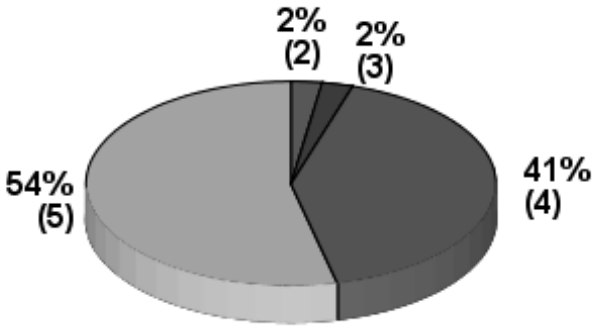
Once the student had ended the monologue, the interviewer had 4 follow-up questions which were designed to widen the scope of the discussion looking at related issues connected with the student and his/ her school, national and international problems. To maintain a standard procedure, the interviewer was instructed to use all the questions, make no changes in the wording of those questions, ask them in the order that they were posed, skipping only those that the student in his/ her monologue had already covered. As can be seen from the chart above, 33 per cent of the teachers found them absolutely and 51 per cent mostly helpful. Thus the teachers mostly appreciated not having to invent questions themselves during the interview, and being able to rely on pre-prepared questions. There is a very small percentage – 4% – of respondents who say that they are mostly not helpful, whereas 12 per cent cannot decide one way or the other. It may be that the final two groups represent interviewers who prefer a freer interviewing style where the follow-up questions are more tied to what the candidate had previously said in his/ her monologue.

Figure 31. Q. 31. The follow-up questions were appropriate. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



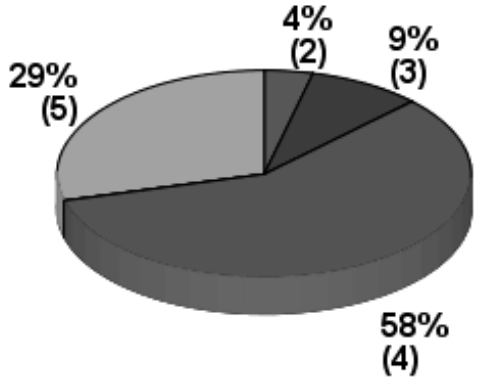
Here, too, most of the respondents, commented positively (30% absolutely, 54% mostly) on the appropriateness of the questions, which the research interprets as well – connected with the monologue topic, and suitable for the age-group at hand to discuss. Only a very small minority considered the questions either absolutely (1%) or mostly (1%) inappropriate for the context. No comments were added.

Figure 32. Q. 32. Role-play is a good task-type for the speaking exam. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



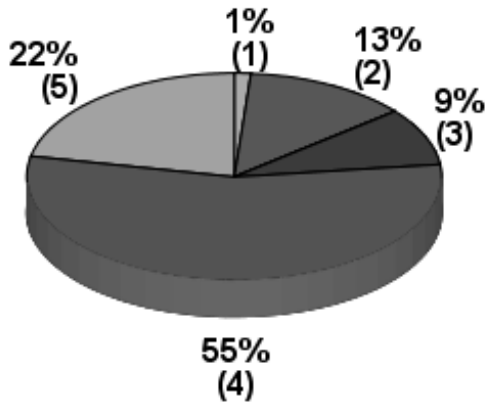
Role-play as the second task during the national examination speaking test has a long history and is thus a familiar and well practiced task type for the teachers – interviewers. This may account for the fact that nearly all participants approve of the role-play as an exam task (54% absolutely, 41% mostly). Only 2 per cent of the respondents consider it mostly not a good task type for the current purpose and another 2 per cent do not have an opinion on the subject.

Figure 33. Q. 33. The topics for the role-play are appropriate. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



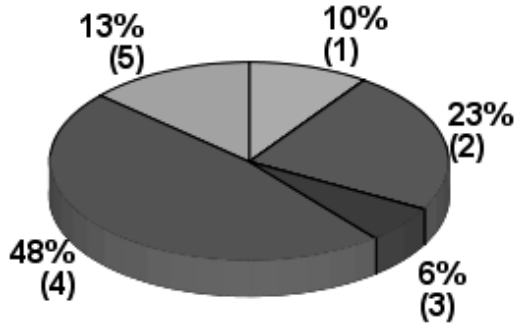
Role-play topics are chosen so that they would represent potential real-life situations that the students should be able to handle if they have reached B2 language level. They are also drawn up based on the topics outlined in the national examination specifications. The teachers who participated in the survey seemed to find the topics suitable (29% absolutely, 58% mostly). Four per cent consider the role-play topics mostly inappropriate and 9 per cent do not think they are able to judge the appropriateness.

Figure 34. Q. 34. I use the exact wording, answering students' questions in role-play. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



Contrary to what the interviewers are expected to do while administering the monologue, during the role-play, the interviewer becomes a participant in the conversation and is thus required to participate in it in a natural way. To facilitate the conversation, the interviewer card contains information that he/she could potentially use while responding to the student's question. It has become evident over the years that interviewers handle that part of the interview very differently, some using it as a prompt to build his/ her answer on, and others as a ready answer to use when the student asks a question, thus risking sounding artificial, irrelevant or inappropriate. The figure above reflects the participants' choices. Twenty-two per cent of the respondents claim never to deviate from the given wording given on the role card and a further 55 per cent say they mostly do not change it. Only 1 per cent of the interviewers claim they never adhere to the given sentence and 13 per cent say that they mostly do not do that. Looking at this data, it seems that the vast majority of the teachers are not aware of the different roles they play as interviewers during the national examination speaking test and further training is needed to standardise the behaviour.

Figure 35. Q. 35. I try to change the answer depending on the question. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

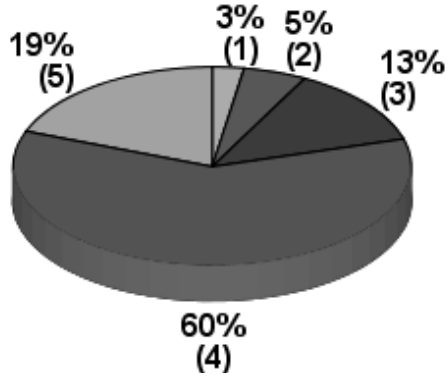


The responses to the statement above are not in accordance with the answers to statement 34 (cf. above). While all in all 14 per cent of the respondents stated changing the wording of the responses in statement 34 (1% + 13%), here 13 per cent of the respondents admit always changing it and a further 48 per cent claim they mostly change it. There is a controversy in the respondents' claims and a study analysing the actual recordings would probably shed more light as to the teachers' actual behaviour during the interviews.

3. 3. 6. The Marking Scale

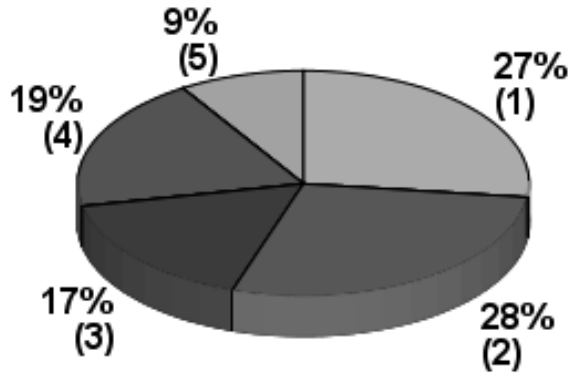
As the goal of the questionnaire was to study the interviewers' attitude towards the newly introduced examination procedure, relatively less attention was paid to the evaluation of the marking scale. Still questions were asked as to its accessibility and ease of use. Statements 36 and 37 asked the respondents to evaluate that.

Figure 36. Q. 36. It is easy to use the marking scale for speaking. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



The data in the graph above suggest that the majority of the respondents find the new marking scale quite accessible (19% absolutely, 60 % mostly). There are 8 per cent of the respondents who have encountered a level of difficulty while doing that (3% absolutely, 5% mostly) and 13 per cent of the respondents find it hard to express an opinion here. It is somewhat surprising to find such a high percentage of teachers who report ease of use, as the new scale differs notably from the earlier one. Also, it is important to further research the points of difficulty in order to either hone the marking scale or provide more training for the assessors in its use.

Figure 37. Q. 37. I need more practice with the marking scale. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

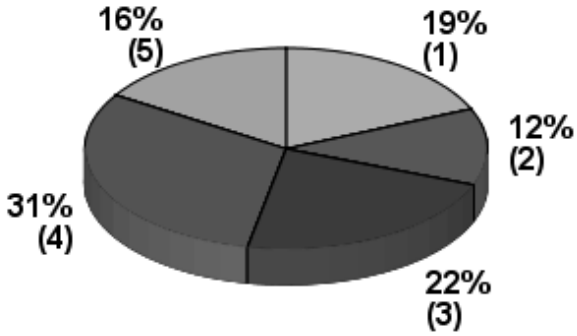


This statement yielded quite a wide spread. Twenty-seven per cent of the interviewers expressed strong confidence in their ability to use the marking scale, saying they absolutely did not need any further training. A further 28 per cent reported this being mostly the case. Twenty- eight per cent of the teachers expressed a need for additional practice (9% absolutely, 19% mostly) to increase their confidence level. Seventeen per cent of the respondents found it hard to respond. For a more objective answer to the question, an analysis needs to be carried out as to how effective the teachers really were in their rating scale implementation.

3. 3. 7. Recording the Interview

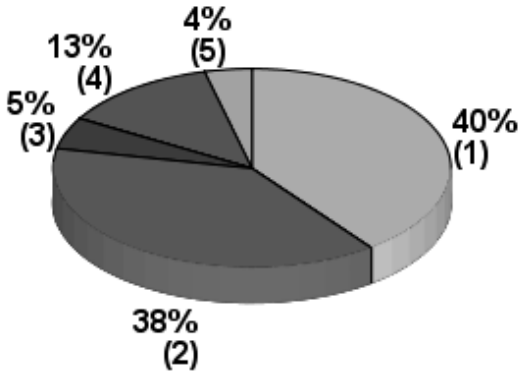
Statements 38 and 39 studied the interviewers' reaction to recording the interview.

Figure 38. Q. 38. I get nervous when the interview is recorded. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



Most people experience a level of nervousness when their performance as an interviewer or an interviewee is recorded. As, from the point of test score reliability, it is imperative that the interview be recorded (to provide a second rating if necessary and to monitor inter-rater and intra-rater reliability), the teachers' attitude to recording was sought. The figure above reflects the teachers' self – assessment on the issue of nervousness while the interview is being recorded: nearly a half of the teachers report some level of nervousness (16% absolutely true, 31% mostly true, more than a quarter denies being nervous during the interview (19 % absolutely, 12% mostly) and 22 per cent of the respondents find it hard to evaluate, which may be an indication that they have never attempted or been required to record student examination interviews.

Figure 39. Q. 39. I record student interviews in class. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.

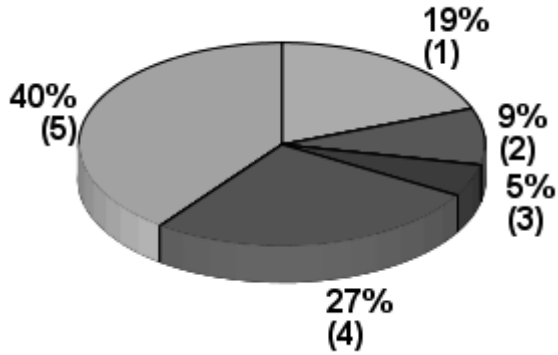


It can be seen from the data above that the overwhelming majority have not recorded student interviews in class (40% absolutely, 38% mostly). A further 5 per cent refrain from reporting on their practices and only 17 per cent of the respondents claim recording them (4% absolutely, 13% mostly). From the point of score reliability it is advisable to change those practices. Also, nervousness during the interview concerning the recording of it may be reduced if recording has become part and parcel of the day-to-day language learning/teaching/assessment practices. A further consideration here is that Estonian students will be disadvantaged at other international examinations, the speaking sections of which are invariably recorded, if they are deprived of the opportunity to practice being recorded in the classroom situation at home.

3. 3. 8. The Examination Room

Providing uniform conditions for examinees also involves providing all examinees with a quiet exam room devoid of disruptive external noise (traffic, construction work, other students during recess, etc). Random checks of earlier recording demonstrate that oral examinations are sometimes held in classrooms that are surrounded by notable noise levels. Thus, a statement was included in the survey regarding the teachers' freedom of choice regarding the room where to conduct the interviews.

Figure 40. Q. 40. I can choose the classroom for the speaking exam. Responses: 1 – not at all true, 2 – mostly not true, 3 – hard to say, 4 – mostly true and 5 – absolutely true. N=81.



The majority of the teachers claim that they have a choice in the matter (40% absolutely, 27% mostly). A worrying 28% have little choice in the matter (19% absolutely, 9% mostly), which may mean that the interviews may be conducted in an environment which is not conducive to spoken language testing (a noisy background, sitting far from the recorder, etc.). Five per cent of the respondents found

it hard to answer the questions, which may mean that either they have not thought about the physical conditions of the interview or that they have never attempted to require a more suitable room for the interviews.

3. 3. 9. Comments to the Questionnaire

In addition to responding to particular statements, survey participants were given an opportunity to comment on pre-examination training and materials, the exam procedure, the marking scale and any other aspect they felt inclined to. Out of the complete figure of 84 participants, 29 used the opportunity. Their remarks ranged from a very laconic note on a particular aspect (e.g. when asked to comment on the exam procedure, the comment was: *correct*, similarly the marking scale earned a comment: *too detailed* or *is good*) which sometimes lacked clarity and were thus hard to interpret (e.g. any other aspect of the exam – comment: *time limit* with nothing added, any other aspect of the exam – comment: *wide, more age appropriate* without further elaboration), to detailed comments on all the aspects listed above.

The comments about training and materials focused predominantly on their quantity and availability of used of (i.e. previous years') examination materials to schools as training material. 13 participants commented on the section. The teachers were of the opinion that the training materials (sample topics for the student monologues) were too few and more should be available on the website. Suggestions were made to make the previous years' monologue topics and role-plays available to schools once the examination results had been announced, so that they could be used in the classroom. There were also complaints about the training video, which was seen to contain 'mistakes' (deviations from the script on the part of the interviewer) and thus be of poor quality. As the same video is used as a sample and a training tool during the pre-examination training, it does pose a problem. As a training tool, having 'mistakes' in it cannot be construed as a flaw, as part of the training is for the participants to find the problems. As a sample on the web-site, it should not contain the flaws any more, because the posted video could potentially be used as a device for independent training and should thus serve as an example of proper interview behaviour. If it is used unchanged on the website, it should at least be supplemented with comments about the interviewer's behaviour.

Comments about the examination procedure (9) seemed to reluctantly favour the suggested paradigm (e.g. *easy to follow thanks to the papers, we are used to it*, *clear and understandable, more or less OK*). There was one comment indicating that recording made the interview very hard, without specifying why this happened. One comment concerned the external candidates (students from other schools to be tested at that particular school) and requested the data about them to be sent to schools earlier than the examination day.

The marking scale was discussed in quite controversial terms. There were assessors who found it accessible and easy to use (e.g. *is good, definitely correct*) and those

who thought the scale could be more detailed (e.g. *could be even more specific, could be even more detailed*). On the other hand, some comments testified to the fact that changing the marking scale is connected with difficulties of adjustment to the new one (e.g. *former scale was easier to use, too complicated to follow*), and old habits may linger in spite of the awareness of the new scale requirements and training (e.g. 'points for monologue and role-play should be different on the scale'). The comments also revealed some of the main difficulties while using the new scale, namely distinguishing particular criteria from one another and making distinctions between criterial levels within a particular criterion (e.g. *the first and the last column in the scale confusing*). There were individual raters who missed having the number of mistakes pointed out within each scale level (e.g. *the number of mistakes would help*). Overall, the need for more experience using the new scale was pointed out, which all the above comments suggest as well.

The any other aspect of the speaking exam section allowed the teachers to voice quite a few other concerns. Some of these comments were procedural and pertained to the time-frame (difficult to observe the time, more flexibility should be allowed from one interview to the next) and the formality of the procedure (e.g. *I should have the right to express my emotions, the warm-up part is too formal*). Certain comments revealed the interviewers uncertainty of the right course of action if the students deviated unexpectedly from the prescribed role during the role-play. They also expressed concern about a possible change in the role-play format, fearing a set-up where the students are evaluated through a peer interview.

3. 3. 10. Results of Cluster Analysis and Correlation Analysis

Cluster analysis and Spearman rank correlation were used in order to establish further patterns among the respondents' attitudes, behaviours and perceptions.

Cluster analysis aimed at revealing significant dimensions that seemed to divide the respondents into dispositional groups. Cluster analysis was thus used similarly to factor analysis. Clusters were assembled based on variables, i.e. clusters comprised of statements that received similar responses from particular respondents. The method of hierarchical clustering was used. Cluster dendrogram can be viewed in appendix 4.2. The discussion below is a content interpretation of the patterns that emerged as a result of cluster analysis paired with Spearman rank correlation for particular statement pairs.

Cluster analysis, then, divided the statements into groups that seemed to draw similar responses from the respondents, i.e. the respondents that marked the higher end on the scale for a particular statement in the cluster tended to do that for also the other statements within that cluster and vice versa, those who marked the low end on the scale tended to do that for the other statements in the cluster. Each cluster was subsequently described as characterising the respondents who had marked the high end of the scale for the statement in that particular cluster.

Cluster analysis appeared to divide the statements into two broad groups, whereas within the second group, two subgroups seemed to stand out. Group one comprised the responses to statements 4, 5, 9, 10, 11, 12, 15, 16, 18, 19, 22, 24 and 37 (cf. appendix 4.2). These statements tackled the need of further training and practice in the use of the examination materials, non-adherence to the script, challenges with time management and the candidates' uncertainty about what was required of them. The correlation strength between particular statements and its significance varied within the cluster. There were strong correlations between responses to statements 4 and 5 ($\rho = 0.590$, $p \leq 0.01$) and 4 and 37 ($\rho = 0.668$, $p \leq 0.01$), indicating the respondents' need for more training and their willingness to take part in the training, which may indicate that the group felt uncertain about their role as a language tester or that they were unclear about either the whole examination procedure or particular aspects of it. There was also a noteworthy correlation between statements 9 and 10 ($\rho = 0, 327$, $p \leq 0.01$) and 9 and 11 ($\rho = 0, 297$, $p \leq 0.01$), which seems to demonstrate the respondents' lack of awareness of the need for a uniform procedure for the speaking test, as those respondents report changing and ignoring the script for various reasons. There seems to be a need for additional clarifications for that group of the difference between the language teaching and a language testing situation. A strong correlation also manifested itself between the responses to statements 18 and 19 ($\rho = 0, 503$, $p \leq 0.01$), where the teachers report students to need explanations and clarifications. Not knowing the nature of those queries, one can but speculate that those teachers who had prepared the students well about the examination procedure or conducted the interview with confidence would get few questions, whereas in a situation where the teacher him/herself was hesitant, more questions would arise. There was also a notable correlation between the responses to statements 16 and 22 ($\rho = 0, 224$, $p \leq 0.05$), pertaining to the respondents perception that the candidates needed more time than was available according to the script, in which case they seemed to award the candidate the time required. The time-keeping itself in the group seem to be a challenge, as indicated by the responses to statements 12 and 15 ($\rho = 0, 20$). It is interesting to note that there is a noticeable correlation between responses to statements 10 and 24 ($\rho = 0{,}290$, $p \leq 0.01$) and 11 and 24 ($\rho = 0{,}321$, $p \leq 0.01$) both of which have to do with the fact that the respondents claim that they either change or ignore the script one the one hand, and that students finish the monologue before their given time, i.e. may not have demonstrated their speaking skill to the best of their ability. These responses may be connected in that the students short responses may have stemmed from problematic input on the part of the interviewer, they did not say enough to prompt the students to give a fuller response or may have prompted the students not to speak very much (with the purpose of avoiding mistakes, for example). The correlation coefficients reported above illustrate instances where strong significant correlations manifest themselves. Ties between other responses within that cluster are not marked by the same amount of strength or significance. To sum up, group one seemed to be ambiguous or uncertain about their role as an assessor/ interviewer and also lack rigor about their own examination procedure. The respondents reported forgetting about time during the interview

and letting the students talk for as long as they wanted. They claimed that the wording of the scripts was artificial and that they tended to change the wording of the frame. The group seemed to struggle with timekeeping in general, noting that it required effort and that the students either finished before time or required more time than they were given. The respondents in this cluster stated that they needed more training and practice and that they were willing to take part in further training. This group seemed to find it hard to shake off their role as an accommodating language teacher who was willing to overlook examination procedural requirements to cater for the students' individual peculiarities and to take on the role of an interviewer whose aim should be to create uniform conditions for all examinees during the interview and to perform in a consistent way from one interview to the next.

Responses to the statements in group two seem to characterise their respondents as being generally satisfied with the new examination procedure and more consistent in their behaviour during the speaking test. This group however appears to have two sub-groups. Sub-group one comprises responses to statements 2, 3, 23, 25, 28, 29, 30, 31,32, 33, 34, 36 and 38 (cf. appendix 4.2). The statements here focused on the appropriacy and helpfulness of the examination materials, tasks and topics, candidate adherence to the preparation time allowed and their readiness to produce a speaking sample of expected length, but also unnecessarily meticulous handling of the role-play task and reluctance to have the interview recorded. Here, too, the correlation strength between particular statements and its significance varied within the cluster. Strong correlations exist between 28 and 29 ($\rho = 0,858$, $p \leq 0.01$) where the respondents comment on the appropriacy of monologue topics, between 30 and 31 ($\rho = 0,629$, $p \leq 0.01$), where they praise the follow-up questions and between 32 and 33 ($\rho = 0,642$, $p \leq 0.01$) where appropriacy of topics and role-play as a task are commended. A strong correlation can also be detected between 2 and 3 ($\rho = 0,550$, $p \leq 0.01$), between 2 and 36 ($\rho = 0,472$, $p \leq 0.01$) and between 3 and 36 ($\rho = 0,412$, $p \leq 0.01$) where the respondents comment on the sufficiency of the pre-examination training, usefulness of the materials and accessibility of the marking scale. A high correlation can be found between responses to statements 30 and 36 ($\rho = 0,414$, $p \leq 0.01$), 33 and 36 ($\rho = 0,345$, $p \leq 0.01$) and between 30 and 33 ($\rho = 0,391$, $p \leq 0.01$). All these statements have to do with estimating how well the examination materials have been prepared. The cluster demonstrates a noticeable correlation between the responses to statements 23 and 25 ($\rho = 0,254$, $p \leq 0.05$), which seems to show that the respondents in this group detected few problems with the students utilising all the allotted time for the monologue. All the above responses seem to characterise the group as being content with all aspects of the procedure: pre-exam training, materials, tasks, topics and the marking scale. These seem to be the conscientious teachers who get on with the task at hand and offer little criticism of the different aspects of the examination. There was an interesting connection within this cluster between responses to statements 34 and 38, which was noteworthy ($\rho = 0,210$), and seemed to indicate a certain level of anxiety. The responses seemed to indicate that recording the interview makes them nervous, which may signal the fear of being caught to have made a mistake (either in language use

or in the examination procedure). It may be the same fear of making a mistake that prompts these respondents to follow the role-play interviewer card verbatim without adapting its content or language to the actual candidate question during the role-play..

Subgroup two comprised of responses to statements 1, 6, 7, 8, 13, 14, 17, 20, 21, 26, 27, 35, 39 and 40 (cf. appendix 4.2). These statements reflect the ease of examination materials use and the respondents' own initiative in seeking information regarding examination material updates as well as providing appropriate conditions for the speaking test. As in the above clusters, the correlation strength between particular statements and its significance varied within the cluster as well. Here strong correlations were found between responses to statements 7 and 8 ($\rho = 0, 435$, $p \leq 0.01$) where the respondents comment on the usefulness of the script on the one hand and ease of its use. The same can be said about responses to statements 13 and 14 ($\rho = 0, 345$, $p \leq 0.01$) where the respondents claimed to keeping to the time frame and not hesitating to stop the students when the time was up. The responses to statements 26 and 27 ($\rho = 0, 565$, $p \leq 0.01$) manifest ease of responding to the questions on the part of the students, which is corroborated by the correlation between 17 and 26 ($\rho = 0, 473$, $p \leq 0.01$). There is a strong correlation between the responses to statements 1 and 40 ($\rho = 0, 303$, $p \leq 0.01$), which may indicate that the respondents here are clear about the requirements set to them as interviewers/ assessors and make conscious decisions about meeting those requirements. There are responses to other statements that appear in the dendrogram in the cluster under discussion (e.g. 35 and 39) that seem to point to the respondents here being decisive and confident about different aspects of the speaking test, but the correlation with the aforementioned characteristics is weak and statistically not significant. All in all, subgroup two can be viewed as the more analytical and also more proactive. They found it easy to keep time during the interview, kept to the required time-frame and the exact wording of the monologue script and could easily stop the students when the time was up. This subgroup also commented on their students responding to the interview task with relative ease: their students understood what they had to do, found the monologue topics easily understandable and thought the questions easy to answer. These respondents checked the examination centre website frequently for new materials, they claimed to be able to choose the classroom for the interview and recorded student interviews in class.

If predictions were to be ventured about the reliability of the above-mentioned groups' evaluation results, then the interview results awarded by the members of the first group would probably need to be second-marked to ensure consistent evaluation. At least procedurally, the second group would provide more uniform conditions.

3. 4. CONCLUSION

To sum up, a number of steps have been taken to provide uniform testing conditions for the candidates during the national examination speaking test: the provision of interviewer scripts along with interviewer behaviour guidelines and a marking scale for the assessors are necessary pre-conditions for obtaining valid examination results.

It is the interviewer behaviour during the oral proficiency interview, or rather, at this point, the interviewer perception of what is expected of him/her during that interview that may potentially lead to differences in the interviewing practices and, consequently, examination results.

All in all, the questionnaire study yielded quite a revealing picture of the interviewer perceptions of the national examination speaking test and their own skills to conduct oral interviews in such a manner that they would yield comparable results. The teachers' commentary in their questionnaire responses testifies to the fact that teachers are very involved in the evaluation procedure and worry about the good practice during the evaluation process. Cluster and correlation analysis shows, however, that they do not see their role in the interviewing procedure in a uniform manner. It is also quite evident that further experience and training is necessary to increase the teachers' confidence level as interviewers and evaluators. They need to know what the expectations are to a fair and professional oral proficiency interviewer and in what way the interviewer behaviour has been known to deviate. They need opportunities to put their skills to the test and they need peer and supervisor feedback on what their own interviewer behaviour is like. On the other hand, once awareness-building and practical conducting of interviews training has been provided, interviewers need to be monitored with regard to the implementation to the required practices. Should it appear that the trained interviewer markedly deviates from the script during the national examination even after training, such interviewers should be excluded from the interviewing practices, as their practices result in diminishing the validity of the examination results.

4. INTERVIEWER BEHAVIOUR DURING THE SPEAKING TEST OF THE ESTONIAN NATIONAL EXAMINATION IN THE ENGLISH LANGUAGE

The purpose of this chapter is to present and discuss the results of a study which was carried out following the 2008 national examination in the English language in Estonia with the aim of investigating interviewer behaviour during its speaking test with a view of discovering if and to what extent the interviewer behaviour conforms to the scripts the interviewers were provided with, and if deviations can be detected, if the latter display any patters of interviewer language that might potentially affect candidate score or national test development.

Fulcher and Davidson (2007) draw the test developers' attention to construct-irrelevant variance (2007:25), i.e. variance in tests results that might originate not from the test taker's control of the construct but from the test – taking context. They define the context as 'the room where the learners will sit, the proctor or invigilator who shows them into their seats and supervises the test, the decoration, temperature, and all the other factors that might impact on the test performance of a person taking the test' (ibid). In the context of a speaking test, this will also include the behaviour of interlocutors and raters. McNamara (2000) emphasises the need to go beyond investigating the candidate while trying to evaluate his/her language proficiency. He urges research to consider 'those who frame the opportunity for performance at the design stage; those with whom the candidate interacts; those who rate the performance; and those responsible for designing and managing the rating procedure. Instead of focusing on the candidate in isolation, the candidate's performance needs to be seen and evaluated as part of a joint construction by a number of participants, including interlocutors, test designers, and raters.' (2000:21).

Although rater behaviour has long been the subject of testing related research (cf. Lado 1961, Bachman 1990, Alderson et al 1995, Fulcher 2003 among others), the investigation of interlocutors is a much more recent development as was shown in chapter one. Fulcher and Davidson (2007) attribute it to the fact that 'we are now much more aware of the fact that discourse is co-constructed, and so the performance of the test taker is in part dependent on the performance of the interlocutor' (2007:132).

Being aware of the interviewer's language variance possibly affecting the test taker's performance during the national examination in the English language in Estonia, the test developers have resolved to provide the interviewers with script, as discussed in the previous chapter. In addition to that, all interviewers were required to participate in the interviewer and assessor training as a precondition to acting either as interviewers or assessors during the national examination.

The current study was set up to investigate to what extent the interviewers' behaviour corresponded to the behavioural patterns that were envisaged for

the interviewers in the national examination interviewer scripts. In order to do that, 50 interviews were randomly selected, transcribed and analysed for the presence or absence of the required interview elements and any other peculiarities that featured on the recordings. The sample for the study was guided by the following principles: the interviews had to have taken place in the spring 2008 national examination session, i.e. they would have followed the new speaking examination framework; there was to be an equal number of interviews both from Estonian and Russian medium schools; other than that there was a conscious attempt to make sure that all the interviews would come from as different parts of Estonia as possible. It has to be admitted that the idea of convenience sampling cannot be completely ruled out as choices could only be made from among the interviews that were recorded not all the interviews that took place during that year's examination session. As was stated earlier (p.64), of the 9293 interviews conducted, only 624 were recorded. There has been no research done into what motivates the students to choose to be recorded or refuse it (although the recording of the interviews is encouraged as a means for the students to appeal the result should he/she feel the need for that), so any speculation about of the characteristics of the student/interviewer body that makes up the recorded interviews' sample is virtually impossible. Thus the choice of the interviews for analysis had to be made from a limited pool. Therefore, also the findings, though interesting, are tentative and should be handled with the above situation in mind.

The language of instruction at school was chosen as a sampling unit because Estonia is in a very favourable position to investigate if interviewer behaviour (and indeed other phenomena pertaining language testing) is in any manner culturally determined. We have 2 groups of schools that follow the same curriculum and function within the same legislation, yet use a different language in operation. Research literature suggests that there may be features of interviewer behaviour that are determined by interviewer gender (cf. chapter 1). It seems intriguing to discover if any of the interviewer behaviours would only appear in either Estonian or Russian speaking schools. Knowing how interviewers from different cultural backgrounds behave in the interview situations is important information from the point of view of interviewer training, so in-service training sessions could be developed to raise awareness of the idiosyncrasies to promote standard behaviour, which is a key component in the quest for validity.

The interviews were all transcribed by the researcher and subjected to a contrastive analysis regarding the degree of adherence to the interviewer scripts on the one hand and interviewer guidelines of general conduct during the national examination oral interviews, which all the interviewers had been familiarised with during training, on the other. The aforementioned documents are publicly available on the NEQC homepage. They are also sent to schools on the examination day and the interviewers are required to re- familiarise themselves with them an hour before the interviews begin.

The 50 recordings were analysed from the following aspects:

1. participant characteristics,
2. recording quality and details,
3. overall interview time,
4. interviewer language during interview introduction,
5. interviewer language during the lead-in to task 1,
6. monologue preparation time,
7. interviewer language during transition from preparation to monologue,
8. monologue management,
9. interviewer language during the lead-in to task 2,
10. role-play preparation time,
11. interviewer language during transition from preparation to role-play,
12. role-play management,
13. interviewer language during closing the interview,
14. other observations.

After the interviews had been transcribed, they were assigned letter-codes and numbers in the random order, interviews E1 to E25 in Estonian – medium schools and R1 to R25 in Russian – medium schools. While providing evidence for the findings discussed below, these code numbers will be used to refer to particular interviews. For the sake of convenience, in the discussion below the two groups – Estonian medium schools and Russian medium schools are referred to as school-types and called Estonian and Russian schools, for short.

Although the data collected within this study will predominantly be subjected to a qualitative data analysis, this will be supplemented by quantitative methods to reveal further salient characteristics of rater behaviour. Statistical data analysis was conducted relying on SPSS for Windows 16 and Microsoft Excel 2007. In order to detect statistically significant interrelations between the interviewer behaviour and the school-type the interviewer was representing or the interviewer gender, Chi-square test was utilised. Also, t-test was employed to study the relationship between interview duration, the interviewer and the school-type.

4. 1. PARTICIPANT CHARACTERISTICS

The participants in the study included 50 interviewers and 50 candidates selected randomly from among all the interviews recorded during the 2008 national examination speaking section. The table below summarises the distribution of the interviewers by gender in Estonian and Russian schools.

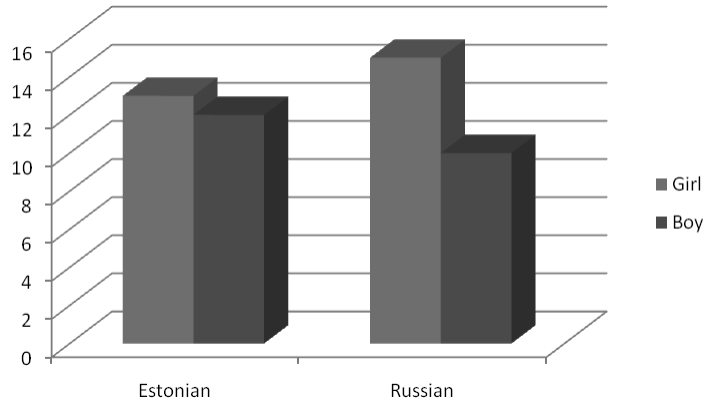
Table 1. Interviewer distribution according to gender and school type.

			Gender		
			Female	Male	Total
School type	Russian	Count	18	7	25
		% within School type	72,0%	28,0%	100,0%
		% within Gender	45,0%	70,0%	50,0%
	Estonian	Count	22	3	25
		% within School type	88,0%	12,0%	100,0%
		% within Gender	55,0%	30,0%	50,0%
Total		Count	40	10	50
		% within School type	80,0%	20,0%	100,0%
		% within Gender	100,0%	100,0%	100,0%

The 50 interviews were conducted by female interviewers on 40 occasions (22 in Estonian and 18 in Russian schools) and by male interviewers on 10 occasions (3 in Estonian and 7 in Russian schools). With the exception of 4 interviews (one in an Estonian school and three in Russian schools), all interviewers were non-native speakers of English. Compared to the interview respondents in the questionnaire study (73 female, 4 male and 5 un-specified), the current population of 80 per cent female and 20 per cent male interviewers will hopefully allow us to investigate interviewer behaviour from the gender perspective more readily and see if the latter plays a part in the process.

Of all the test-takers, 28 were female (13 Estonian and 15 Russian) and 22 were male candidates (12 Estonian and 10 Russian). The table below represents candidate distribution in Estonian and Russian schools according to candidate gender.

Figure 1. Candidate distribution according to gender and school-type (X-axis: school type; y – axis: count of students).



Recording the interview during the speaking test is optional for the student and the candidate body was analysed from the point of gender distribution to see if there was any gender bias in the selection. The above table demonstrate a slight difference between Estonian and Russian schools: girls made up 60 per cent of the candidates in Russian schools and boys constituted 40 per cent of the candidate body. In the Estonian schools, the respective numbers were 52 per cent and 48 per cent, which seems to demonstrate a slightly more even divide between genders in Estonian schools than in Russian ones. The student/ candidate related data will only be discussed in the current study insofar as it contributes to the discussion of the interviewer behaviour. In other respects, it remains outside the scope of the current research.

4. 2. RECORDING QUALITY AND DETAILS

The guidelines to the interviewers for recording the interview (e.g. when the tape-recorder should be switched on and off at the national examination speaking section in the English language, etc.) are quite specific (cf. Interviewers' and Assessors' Procedures Appendix 2). If the candidate requests recording, it should be switched on before the introductory phase and switched off only once the role-play has been completed and the interviewer has announced the completion. With the interviews analysed, on the majority of occasions, the interviewer complied with the requirement: out of 50 interviews, there were 3 (E23, E24, E25) where the recorder was switched on only when the candidate started his/her monologue, and consequently, the interviewer behaviour could only be observed during the last two thirds of the interview (Task 1 and Task 2 but not during the introduction and preparation for Task 1). Similarly, there were 4 occasions (E1, E2, E19, E20) where the recorder was switched off before the end of the interview had been announced. From the point of re-assessment, such practice will cause problems for the second marker. If second assessment was requested for those performances, the assessor had to make his/her decisions on partial evidence and thus possibly misjudged the quality of the candidate's performance.

Recording quality, contrary to what might have been anticipated, was sufficient to make verbatim transcriptions in the majority of cases, which means that the recorder was set close enough to both the interviewer and the test-taker. Only on three occasions would second marking of the student performance have been impossible because of the recording quality (R15, R16, E17). On two occasions (R15, R16), the recorder was set close enough to the interviewer to evaluate her behaviour but too far from the student, as it was only possible to register that the student was responding to the interviewer and the tasks but the content of it was incomprehensible. On the third occasion (E17), the recorder functioned perfectly until the student had completed the introductory stage and the monologue preparation phase, but disappeared completely, as she started with Task 1. Here, too, reassessment would have been extremely problematic. All the above interviews, however, were included in the analyses of the sections that could be transcribed, as they represent real instances of recordings

that have been sent to the National Examination and Qualification Centre as documents of candidate performance that could be subjected to second or third marking if requested. In this respect, they represent problems that need to be addressed during the examination validation process.

4. 3. THE OVERALL INTERVIEW TIME

While discussing the framework developed for the speaking section, it was estimated that, provided the interviewers adhered to the script and the candidates responded to all the tasks, the time required to get a rateable sample from each candidate would fall between 14 and 16 minutes (cf. section 3. 2. 1.). During the analysis, the length of every recording was measured even if the interview did not include all the required parts (introduction, Task 1, Task 2). However, in cases where it could be assumed that the particular parts of the interview were conducted but where the interviewer had not switched on the recorder until the student started presenting the results of Task 1 (E28, E29, E30), or when the recorder had stopped recording in the middle of the interview (E17), the overall time of the interview was excluded from further calculations and analysis that required the whole interview.

Thus of the 50 interviews, 46 could be further analysed for time generally spent on an individual interview and any deviations from that time. Features that transpired in the analysis of the time spent by each interviewer on the whole interview can be summarised as follows.

Table 2. Overall interview time.

N	Valid	46
	Missing	4
Mean		12 min 7 s
Std. Deviation		1 min 30 s
Range		6 min 4 s
Minimum		8 min 32 s
Maximum		14 min 36 s

The average time spent on the oral interview was 12 minutes 7 seconds, the shortest interview (still including all the parts) lasted 8 minutes and 32 seconds (R18), the longest 14 minutes and 36 seconds (R23). The table below demonstrates the temporal differences depending in different school-types.

Table 3. a. Interview duration in different school types.

School-Type	Longest Interview	Shortest Interview
Estonian schools	14 minutes 32 seconds	9 minutes 48 seconds
Russian schools	14 minutes 36 seconds	8 minutes 32 seconds

Table 3. b. Interviewer duration variance in different school types.

	N	Mean	Minimum	Maximum	Std. Deviation
Estonian schools	21	12min 0 s	9min 48s	14min 32s	1min 27s
Russian schools	25	12min 13s	8min 32s	14min 36s	1min 33 s

The table 3.b. above shows that the mean time spent on the interview in Estonian schools was slightly shorter than in Russian schools, but only marginally. This, alongside with very similar standard deviations – σ = 1 minute 27 seconds in Estonian schools, σ =1 minute 33 seconds in Russian schools – leads us to believe that the interviewer length does not depend on the school type. T-test was applied to further corroborate the findings (see appendix 4.4 for t-tests). To begin with, the following hypotheses were established:

H_0 : interview length is <u>not</u> dependent on school type

H_1: interview length is dependent on school type.

The validity of either hypothesis can be decided by comparing the significance established (p) and the required significance level α., set at 0.05. The t- test yielded the result p= 0.6. For the null hypotheses to hold, p> α, which is the case, further substantiating that the interviewer length did not depend on which school type it was carried out.

The interviewer length was further analysed for the difference depending on the gender of the interview. For that the following hypotheses were established:

H_0 : interview length does not depend on the interviewer gender

H_1: interview length depends on the interviewer gender.

The p value of .032, being below α, leads us to conclude that the interviewer length did indeed depend on the interviewer gender (see appendix 4.4 for t-tests.). Further analysis tried to establish gender differences in different school types. Tables 4 and 5 illustrate the findings in Estonian and Russian schools respectively.

Table 4. Interview length and interviewer gender in Estonian schools*.

Gender	Count	Mean	Std. Deviation
Female	19	11min 46s	1min 19s
Male	2	14 min 9s	0 min 32s
Total	21	12min 0s	1min 27s

* 4 interviews could not be analysed for overall time

Table 5. Interview length and interviewer gender in Russian schools.

Gender	Count	Mean	Std.Deviation
Female	18	12min 0s	1min 35 s
Male	7	12min 48s	1min 23 s
Total	25	12 min 13s	1min 33 s

The data above indicate a greater variance in the interview length between men and women in Estonian schools than in Russian schools. Men's interviews compared to their women's counterparts in both schools were longer but noticeably so in Estonian schools – by about 2 minutes. There were but two interviews in Estonian schools that were carried out by male interviewers, so the statistics allow few if any grounds for generalisations. It is noteworthy, however, that both interviews conducted by male interviewers markedly exceeded the mean of the female interviewer time.

While developing the interview format, it was assumed that the temporal variance of the interview would mainly stem from either the tempo of the interviewers speech or the tempo of the candidate's response and the content of it, i.e. a positive correlation could conceivably be expected between the speed of the interviewer's speech and the time spent on the interview. More importantly, though, the time difference was expected to originate from the test-taker's response – how quickly he/she responded, how fast he/ she spoke, how much time he/she would take preparing for the respective sections of the interview (i.e. would he/ she spend all the time allowed for preparation) and how much he/she would have to say. If properly managed, this variance in the candidate behaviour would fall within the 2-minute difference that was allowed by the framework.

As can be seen from the table above, however, the difference between the shortest and the longest interviews far exceeds the 2-minute difference allowed, being 6 minutes and 2 seconds. This may have meant more than a 6-minute advantage for particular students over their peers to demonstrate their language proficiency and conversely, an up to 6-minute disadvantage to others. Another striking observation was that out of 46 interviews only 5 interviews (1 in Estonian schools and 4 in Russian schools) crossed the 14- minute margin, which was deemed minimally sufficient to obtain a rateable sample while still providing all the support necessary for the candidate. On the other hand, just 2 interviewers managed to conclude the interview within less than 10 minutes (E11, R.18). This seems to place the actual time the interviewers required at somewhere between 10 and 14 minutes.

Consequently, questions arise as to how justified the test developers were at setting the timeframe at 14 to 16 minutes. Can a rateable sample be obtained at a shorter time? This question cannot be answered without looking at the amount of student contribution during the interview. The latter, in its turn, is closely connected to the interviewer support provided in terms of adherence to the script. Thus, the questions to be answered concern the time the candidates spent on preparing for Task 1 and Task 2 and the time they spent responding to the task requirements. On the other hand, what this dissertation is more concerned about is the interviewer behaviour at various stages of the interview, congruence between what the script tells the interviewer to say and what the interviewer actually says while guiding the test-taker though the oral proficiency interview. The discussion below will have the abovementioned aim in view while focusing on each of the stages and stage elements separately.

4. 4. INTERVIEWER LANGUAGE DURING THE INTERVIEW INTRODUCTION

In the introduction, the script required the interviewer to

- state the number of the candidate,
- greet the candidate,
- introduce himself (either by stating his/her name and the role he/she was in – the interviewer, or just the role if the student already knew the interviewer's name) and the assessor (both name and the role were required),
- ask the student how he/she was and respond appropriately (cf. Script for the Introduction pp.52–53),
- ask the warm-up questions listed in one of the 5 scenarios provided in the script (cf. Script for the Introduction).

All the interviews were analysed for the presence or absence of the above-mentioned elements, provided the recording contained the section. The latter concern partly excluded interviews E28, E29 and E30 from the analysis (in all the three interviews the candidate numbers were announced). So, in the table below, except for the firsts criterion, where all the 50 interviews were included, criteria 2 to 5 were checked with 22 interviews in Estonian and 25 interviews in Russian schools. There are two figures presented for every element in both columns. The figure before the forward slash gives the number of respective cases in Estonian schools and the number following the forward slash those of Russian schools. The tables 6, 7 and 8 below summarise the findings.

Table 6. Interview introduction element representation in the interviews.

No.	Element	Present in Interviews	Missing in Interviews
1.	Candidate number	24/20	1/5
2.	Greeting the candidate	9/20	13/5
3.	Introducing the interviewer and assessor	9/14	13/11
4.	Inquiring about well-being	13/20	9/5
5.	Asking warm-up questions	22/24	0/1

Table 7. Interview element representation in Estonian schools by gender.

		Female (N=19)	Male (N=3)	Total (N=22)
Introduce	yes	8 (42%)	1 (33%)	9 (41%)
	No	11 (58%)	2 (66%)	13 (59%)
Greetings	yes	8 (42%)	1 (33%)	9 (41%)
	No	11(58%)	2 (66%)	13 (59%)
Well-being	yes	12 (63%)	1 (33%)	13 (59%)
	No	7 (37%)	2 (66%)	9 (41%)

Table 8. Interviewer element representation in Russian schools by gender.

		Female (N=18)	Male (N=7)	Total (N=25)
Introduce	yes	10 (56%)	4 (57%)	14 (56%)
	No	8 (44%)	3 (43%)	11 (44%)
Greetings	yes	16 (89%)	4 (57%)	20 (80%)
	No	2 (1%)	3 (43%)	5 (20%)
Well-being	yes	15 (83%)	5 (71%)	20 (80%)
	No	3 (17%)	2 (29%)	5 (20%)

The tables above reflect some of the dissimilarities between the interviewers with regard to the script. As we can see (cf. Table 6), there is no element that was adhered to by all the interviewers without exception. The first element analysed – stating the candidate number before beginning the interview – showed that the candidate number was skipped by 6 interviewers (E1, R4, R17, R18, R21, R23) and could only be retrieved by checking the label on the sleeve of the tape. Although not affecting the student performance per se, a missing candidate number could potentially lead to errors in score reporting should the sleeve of the tape be lost, damaged or misplaced. This in its turn could undermine the test reliability and hence, validity.

Greeting the candidate also varied from interviewer to interviewer. The respective figures in table 6 above show that the interviewers here fall into two groups, 29 (9/20) interviewers started the interviewer with a greeting and 18 (13/5) did not. Another observation at this point seems to divide the interviewers according to school-type: the greeting in Estonian schools was present much less frequently (9 times out of 22) than in Russian schools (20 times out of 25). Greeting was more frequent with Russian women (89 per cent) than Russian men (57 per cent). The proportion of Estonian women interviewers who greeted the candidate was 42 per cent and that of men 50 per cent. One might speculate here that the reason why greeting was skipped during the recorded interview was because the phase (greeting) had been gone through when the candidate entered the examination room, before the recorder was switched on. When the interview proper was started, i.e. the recorder was switched on, the interviewer thought it unnatural to start with the greeting again. Thus the interviewer starts deviating from the script due to the pragmatics of the situation. Yet he/she ought to be able to project his/her behaviour from the point of view of the second (or third marker) who has nothing but the recording to assess the candidate's linguistic behaviour. A greeting is the opening statement of the script, and, apart from being a common courtesy and a logical starting point for the interaction, it provides structure to the interview, signalling to the candidate that the interview has started. It is an easy turn to respond to, being part of a well-known adjacency pair.

The requirement to introduce oneself and the assessor (cf. Table 6) yielded the results whereby 23 (9/14) interviewers introduced themselves at the beginning of the interview and 24 (13/11) did not. Studying introducing from the point of view of gender, fairly similar results can be observed among female and male interviewers in

Russian schools – greeting was observed with 56 per cent of women and 57 per cent of men. The proportion of women who greeted the candidates in the Estonian schools was smaller – 42 per cent. As with the previous element (greeting) men's proportion (50 per cent) can only be calculated tentatively here, as the number of participants was very small. While designing the script, introduction was incorporated into it in order to establish a professional atmosphere in the room, and to define the roles of the people present – to clarify to whom the candidate was going to be speaking, and who would conduct the actual assessment. The fact that over a half of the interviewers found it unnecessary to go through the introductory phase may stem from the fact that the interviewers were familiar with the students and thought it unnatural to introduce themselves once more. Since this is only a speculation and it is not always the case that the interviewer will interview his/her own students as part of the English language national examination, but will frequently need to interview students he/she has not taught either from his/ her own school or external students from other schools, introducing oneself and the assessor would establish initial rapport between the interviewer and the candidate and deflect some of the anxiety related to speaking to a virtual stranger at the start the oral interview.

The latter issue – removing some of the examination-related anxiety – was at the heart of integrating the query *How are you today?* in the introduction. It was intended as a chance for the candidates to confirm to the interviewer that he/she is managing fine, or voice, and thus automatically slightly relieve, some of his/her anxiety and give the interviewer a chance to further reduce the anxiety-level. With the interviews analysed, 33 interviewers (13/20) recognised the need and 14 (9/5) skipped that section of the script. Here, too, a distinction can be seen between interviews in Estonian and Russian schools. Although more interviewers include the question in the introduction than skip it in both school-types, the proportion of those interviewers who skip it is bigger in Estonian than in Russian schools – 83 per cent of women and 71 per cent of men in Russian schools asked about the candidate's well-being during the introduction, whereas 63 per cent of women and 50 per cent of men did so in Estonian schools. Possible significance of interviewer gender and school type here were assessed with the help of chi-square. A statistically significant strong difference manifested itself between Estonian and Russian schools in greeting the candidate (phi=0,4, p=0,06). Results can be found in appendix 4.1. Investigation of gender significance with the chi-square-test did not display a statistically significant difference. This does not mean, however, that it would not exist. The question should be further investigated with a research population that would include a statistically considerable amount of men.

Warm-up questions were overwhelmingly deemed necessary by the interviewers, as only one interviewer (R22) proceeded to letting the candidate choose a topic right after stating the candidate number without giving the candidate any warm-up period. The vast majority of the interviewers confined themselves to the questions prompted by the script without developing or amplifying the student answers. Two kinds of deviant behaviour were noted: on two occasions (E4, E5) the interviewer treated the script

questions as an introduction to the warm-up proper and proceeded to ask additional questions stemming from the student answers. E4 added 6 questions to the script and E5 added 4 topic-related questions. (It should be noted here that E5 was the longest interview recorded in the respective school-type). In neither case was there an obvious reason for the interviewer to add to the number of questions proposed. In case of E4, the interviewer may have been prompted by the fact that the candidate expressed reluctance to speak on the topic proposed.

E.g.

In. Ok. Let's talk about sports for a start. What sports have you ever tried?

St. Oh, god. I have tried badminton and nothing else really. I'm not a big sport person…….. and try different things but that doesn't mean that I'm good at it. Yea, I'm not a big sports person. I used to watch basketball but I got tired of that.

In. So you are not a big sports person. (not in script)

St. No. …. I can't talk about it.

In. What do you like to do besides doing sports? (not in script) (E4)

In the above example, the interviewer should have recognised that the candidate had produced quite a long turn already and was consequently quite warmed up for the interview. The interviewer should have proceeded to the next question suggested by the script and not prompted the student to choose her own topic to develop and spend unnecessary time on.

The other type of deviation noted was when the student clearly had problems responding to the interviewer questions (e.g. R12). The interviewer, in this instance, seemed to be at a loss as to how to proceed as the student did not respond to the questions at all. As the guidelines did not make that provision, the interviewer, rather than ending the interview altogether, started to prompt the candidate in the hope of getting the interview off the ground.

E.g.

(1) In. Mhmh. Aaa . Olga, let's talk about cooking. What kind of dishes do you like to prepare?

(2) St. I like.

(3) In. What kind of dishes?

(4) St. aaa …….. (a long silence)

(5) In. Fruit , vegetables, or soups, salads?

(6) St. Fruit, vegetables aaaa salad aaa and soup.

(7) In. Why do many people like cooking?

(8) St. …….. (a long silence) What?

(9) In. Why …… do … many people …. like cooking? (slow, clearly articulated)

(19) St. I think aaaaa ….. this is …. very … interesting … ….. and …

(11) In. This is very interesting. (repeats student answer)

(12) St. Mhmh.

(13) In. Thank you. Olga.

In addition to prompting possible answers, the interviewer also resorts to considerably slowing down her speech while repeating the question at the candidate's request. This tactic is probably also chosen to facilitate comprehension between the interviewer and the candidate, and, as will be demonstrated in the discussion below, is something that interviewers frequently resort to at various stages of the interview. The excerpt above also illustrates a strategy sometimes resorted to by the interviewers to increase the overall interview tempo: the interviewer ends the candidate's turn when he/ she realises the candidate is either not able to do so himself/herself (i.e. has nothing to say), or decides to prevent the student from letting to proceed with his/her turn (possibly for fear that the candidate may generate substandard language). In the above excerpt, in line 11, the interviewer, rather than let the student go on with her reply, summarises the candidate's previous response with a falling intonation, thus signalling that the turn can be considered completed to which the student readily agrees by using the agreement marker 'mhmh'.

4. 5. INTERVIEWER LANGUAGE DURING THE LEAD-IN TO TASK 1

Once the introduction had been completed, the interviewers were required to lead the student to the preparation phase of the first rated task – the monologue – quoting the script verbatim. The lead-in to Task 1 was worded so that it would tell the candidate exactly what was expected of him/her during that phase.

> **Interviewer**: Now, I would like you to speak on a topic for two minutes.
> Before you talk, you have 3 minutes to think about what you are going to say. You can make some notes if you wish. Do you understand?
> Here is a pencil and some paper. Please, pick a topic.
> What's the number of your topic?
> Now you will have three minutes.

According to the lead-in, the interviewer has to inform the candidate of the following:

1. essence of the task – speaking on a topic,
2. expected time (length) of the monologue,
3. a 3-minute preparation opportunity,
4. the option of making notes.

The interviewer should also

1. enquire about comprehension,
2. provide pen and paper,
3. request selection of topic,
4. ask the student the topic number,
5. state the beginning of the 3-minute preparation phase.

An attempt was made during the design of the lead-in to arrange the information in it so that it would provide maximum information for the candidate in a concise and logical manner. The information was presented in short, predominantly simple sentences in language, well below the level expected of students on B2 level on the CEFR scale.

The interview tape-scripts were analysed by comparing and contrasting the sentences in the actual interviewer turns found in the recordings with those prescribed in the script. To the elements listed above, three other points of analysis were added. First, in addition to checking for the presence or absence of the required information, research tried to see if the interviewers kept to the order of the information proposed. This was done in order to discover if any patterns would emerge if the order was changed. This, in its turn could mean that a change in the order of information presentation could be made in the scripts, if the change in the sequence would prove more logical and helpful for the candidate. A further point of query was about the wording of the lead-in information. This query, too, attempted to justify the choice made for the current wording by looking at how readily the interviewers utilised it, and if not, what sort of changes were made. Connected to the latter, the third additional point of analysis was about additional information the interviewers included in the lead-in while giving instructions to the candidate.

The results are summarised in the table below. The number of interviews included in the current analysis was 22 in Estonian schools (three recordings started with the student monologue and could therefore not be analysed) and 25 in Russian schools. The figures for the Estonian schools are presented on the left of the forward slash and for the Russian schools on the right side of the forward slash.

Table 9. Element representation in the lead-in.

No.	Element	Present in Lead – in	Missing in Lead-in
1.	Essence of the task	18/23	4/2
2.	Expected time (length) of the monologue	19/22	3/3
3.	A 3-minute preparation time	22/25	0/0
4.	The option of making notes	20/22	2/3
5.	Enquire about comprehension	14/16	8/9
6.	Provide pen and paper	20/23	2/2
7.	Request selection of topic	22/25	0/0
8.	Ask the student the topic number	22/25	0/0
9.	State the beginning of the 3-minute preparation phase	14/9	8/16
10.	Order changed	6/9	16/16
11.	Wording changed	10/20	12/5
12.	Additional information	7/12	15/13

The data above seem to demonstrate that of the nine required elements in the script, there were three elements that were found in all interviews: all interviewers noted the 3-minute opportunity to prepare for the monologue, they did ask the students to select the topic and they did ask the student the topic number. Those elements seemed to be minimally necessary to involve the student in Task 1 preparation.

E. g. In. Ok, You need to choose a topic.
St. Ok.
In. And what's the number, please.
St. aaaa B 3.1.
In. And starting from now you have three minutes, take your time and then you can talk. (E5)

E. g. In. So your number is XXXXXX. Ok. You can choose. ….. So what is the number.
St. Number A 3.1
In. Good. So now you have three minutes to prepare. I'll tell you when the time is up. (R22)

In both cases, the interviewer has ignored the proposed script and created his/her own, retaining only the elements which he/she has found important. Interestingly, both interviewers have added information that is not present in the script. In both cases, the interviewer has considered it necessary to advise the student not to think about time ('take your time', 'I'll tell you when the time is up'). This has possibly been done with the aim of eliminating anxiety during preparation.

Every other lead-in element listed above occurred in the interviews in varying degrees, occurring in a variety of sequences and being expressed relying on a variety of wordings.

a) Essence of the Task

The number of interviewers (out of a total of 47) who informed the candidate what the up-coming task was going to be was 41 (87.2 %) and those who did not inform the interviewer was 6 (12.8 %). There is no gender distinction here, as on three occasions the interviewer was male and on three occasions a female.

b) Expected Length of the Monologue

The proportion of the interviewers who thought it necessary to remind the student about the time he/she was going to be expected to be speaking compared to those who did not do so was identical to that of the essence of the task (87.2% vs. 12.8%). The gender distinction is slightly to the disadvantage of male interviewers, as 4 times out of 6 it was the male interviewer who did not mention the expected length of the monologue.

c) The Option of Making Notes

Here, too, most of the interviewers adhered to the script (42 out of 47, i.e. 89.4%), and a small minority (5 interviewers, 10.6%) refrained from doing so explicitly. Of those, three (E1, R3, R4) overtly referred to the opportunity by telling the student further in the lead-in that they had pen and paper at their disposal. As all the interviews analysed contained a period of time when the candidate seemed to be preparing for the monologue in one form or another, the interviewers who did not mention it in the lead-in may have worked with a belief that, as students had been prepared with regard to the interview procedure ahead of time, there was no need to reiterate that. The interviewer should have remembered, however, that he/she may have interviewed candidates who had been introduced to the interview by somebody who may have forgotten or ignored procedural instructions. From the point of view of clarity of the procedure, and given the fact that the candidates are in a situation in which most of them have never been before, and therefore anxious, every effort has to be made to obtain as good a performance from the candidate as possible, which includes explicit rather than implied instructions.

d) Enquiring about Comprehension

Compared to other elements of the lead-in, asking whether the student had understood what he/she was expected to do seemed to have been left out more often than other elements: out of 47 interviewers, 30 kept it in the interview (63.8%) and 17 did not (37.2%). The question 'Do you understand?' was included in the lead-in in order to give the student an opportunity to ask the interviewer to repeat the instructions if the tempo of the interviewer's speech had proved to be too fast or if the candidate had been too anxious to pay attention to the instructions. A question arises what prompted some of the interviewers to overlook the element. It may have been lack of training and subsequent disregard to the function of different elements in the lead-in. It may also have been that as candidates had demonstrated good language ability during the introductory stage, the instructions given in fairly simple language seemed to render such a question superfluous. It is worth noting, perhaps here that in all the interviews analysed, when the question was asked, it never received a negative response.

e) Providing Pen and Paper

The comment regarding pen and paper was incorporated in the script to contribute to clarity of procedure. A great majority of the interviewers 43 out of 47 kept it in the lead-in (91.5%), and there were 4 (8.5%) who did not. In the latter, in two cases pen and paper being provided was implied (E18, R15) and in two cases (E5, R22) it was neither mentioned nor implied. An example of the case where the provision was implied can be seen below.

E.g. In. Ok. But now weeeeee …. Proceed to the second stage. And I'd like you to pick a card. …. So what is the number?
St. A 3.1

In. A 3.1. And now you have three minutes to think about what you're going to say. You can make notes and after that you'll start the monologue and you'll speak on the topic for two minutes.

The interviewer mentions the three-minute preparation time and provision of pen and paper is implied by pointing out the opportunity to take notes. Similarly with other elements of the lead-in, however, all parts of the instructions should be clearly stated to the candidate to provide support to the candidate and structure and clarity to the generally very stressful examination situation.

f) Stating the Beginning of the 3-Minute Preparation Phase

At the start of the lead-in, the interviewer script requires the mention of the 3-minute preparation phase. It was considered necessary, however, to indicate the start of the phase once again at the end of the lead-in. It was thought relevant mainly because a fair amount of information had been conveyed to the candidate after first mentioning the preparation phase (taking notes, understanding instructions, getting pen and paper, choosing a topic, stating the number) and it was important to signal the starting point to the candidate precisely. Marking the beginning was also important from the time-keeping perspective. In the interviews analysed, fewer interviewers retained it in their lead-in (23 out of 47) than dropped it (24 out of 47), but only marginally. Leaving the signal out of the lead-in, however, may not only have contributed to the ambiguousness from the candidate's point of view as to how much time he/she had available, but it may also have had a detrimental effect on the interviewer's own time-keeping.

g) Changes Introduced in the Lead-In

The changes that interviewers introduced in the lead-in can be divided into three groups:

1. Changes in the sequence of information.
2. Changes in the wording of information.
3. Additional information given besides what was included in the script.

Table 12 above shows the number of interviewers who, although keeping the necessary information in the lead-in, presented it in a different order. The number of people who introduced a variation in the **sequence** was 15 (6/9) – 31,9 per cent and of those who did not 32 (16/16) – 68.1 per cent. Analysing the different information sequences that were suggested by interviewers, a pattern emerged: rather than explain the task and then ask the candidate to choose a topic, some interviewers asked the candidate to choose a topic first and then proceeded to outline the task. There were six (3/3) interviewers who made the respective alteration. An example can be seen below.

E.g. In. Ok. But now weeeeee …. Proceed to the second stage. And I'd like you to pick a card. …. So what is the number?

St. A 3.1
In. A 3.1. And now you have three minutes to think about what you're going to say. You can make notes and after that you'll start the monologue and you'll speak on the topic for two minutes. (E18)

In the example above, we can see that the interviewer provides the candidate with most (but not all) of the information in the script, but the request to choose a topic comes much sooner than the script actually proposes. While there is no rule as to what the order of the information in the instructions should be, there is an underlying belief that it should be as logical as possible, and interfere with the preparation per se as little as possible (Tankó 42). Giving the monologue topic to the candidate before the instructions regarding how to handle the task, the interviewers run the risk of not getting the candidate's proper attention regarding the procedural details (time allotted for preparation and speaking, note-taking, etc.) as the candidate will already be concentrating on reading the task card, thinking about the statement and not listening to the interviewer any more.

As far as the **wording** of the lead-in was concerned, 30 interviewers (63.8%) changed the wording of the script in one way or another, whereas 17 interviewers (36.2%) used the scripts verbatim. Here a difference can be observed between the school types. The interviewers in the Russian schools changed the script much less frequently (20%) than their counterparts in the Estonian schools (45.5%).

The changes fall into several categories:

1) using a synonym instead of a term suggested in the script;

E.g. In. Mhmh , ok, thank you. Now I would like you to speak on a topic for two minutes. Before you talk you have three minutes preparation time, so you can take notes, some paper and pen, but ….. I want you to take a topic first. Choose a card. ……. And tell us the number please. (E12)

E.g. In. So, here is a pen and some paper. Please pick a topic. ……. And what is the code number? (E21)

There were 15 (4/11) instances of interviewers who used a synonym rather that the term suggested. In the first example above, the interviewer has used a much vaguer term 'preparation time 'instead of a more concrete 'to think about what you are going to say'. She has also expressed the request to choose a topic two times: instead of 'please, pick a topic, the interviewer first tells the student to 'take a topic', immediately followed by 'choose a card', probably hoping that the student would comprehend at least one of the phrases. In the second example, the interviewer uses the term 'code number' instead of the 'number of your topic'. The use of the term is erroneous as the term 'code number' is used to denote the number allocated to each candidate for the examination period for identification and in this respect may cause confusion.

2) using bald-on-record expressions to give instructions rather than politer forms enclosed in the script;

E. g. Take some paper over there and …pen …and… be prepared in three minutes time (R4).

E. g. Before you talk you have three minutes preparation time, so use the paper, use the pen (E13).

E.g. So pick your topic. (E18)

By using the imperative forms without any mitigation the interviewers appear unnecessarily direct and inadvertently emphasise the unequal power positions in the examination situation and may thus contribute to the stressfulness of the situation. There were 3 (2/1) recorded instances of choosing a bald-on-record expressions in this section of the interview.

3) reducing the level of politeness;

E.g. You can choose any card, please (R12).

E.g. Ok, starting from now you have three minutes to prepare, you can take notes (E4).

E.g. You'll start the monologue (E18).

In the first example above, the interviewer resorts to an unnatural wording of an instruction where she first expresses permission (you can choose any card) and follows it up somewhat illogically by the politeness form (please) usually added to requests. It seems to be a case of the interviewer trying to soften the initially fairly direct turn to the candidate. In the second example, the interviewer has omitted the clause 'if you wish' present in the scrip at the end of the phrase and has retained only the possibility/ permission part of the instruction ('you can take notes'). In the third case, the interviewer has included an additional instruction in the lead – in and has done so in the form of a statement/order rather than a polite request. All the above examples represent cases where the interviewer has lowered the politeness level, where the interaction seems to be taking place between unequal participants, where the interviewer has assumed more power that the scrip originally assigned him/her. There were 7 interviews (all recorded in Estonian schools) where the politeness level was reduced compared to the script.

4) resorting to either grammatically or lexically erroneous structures;

E.g. In. Thank you. Now I would like you to speak on a topic for 2 minutes. Before you talk you have 3 minutes to think about what are you going to say. You can make some notes if you wish. (R4)

E.g. In. Thank you. Now I would like you to speak on a topic for two minutes. Before you talk, you have three minutes for your thinking about what are you going to say by the topic. (R13)

All in all seven interviews (2/5) contained either a grammatical or a lexical error in the lead-in delivered by the interviewer. The first example above illustrates the most common grammatical deviation that occurred in the use of the indirect speech. Here direct speech word order was erroneously introduced. The second example, besides

including an error of the type mentioned already, contains the expression 'by the topic', a non- existent collocation. Other recorded errors occurred in the use of prepositions and in subject-verb agreement. All types of errors could have been avoided if the interviewers had followed the script given to them.

Sometimes, the interviewers thought it necessary to include **additional elements** in the lead-in. This seemed to have served the purpose of contributing to the clarity of the instructions delivered, but occasionally may have resulted in giving the candidates false information.

Above we saw an example of when the interviewer reiterated the request to the candidate to pick a topic. Similar behaviour, i.e. repeating particular elements in the lead-in could be observed with other segments as well.

E.g. In. Ok. Aaa please pick a topic among these. …… show me the number of your topic. A four two…….. You have three minutes. You have three minutes to prepare. (R24)

E.g. In. Mhmh , ok, thank you. Now I would like you to speak on a topic for two minutes. Before you talk you have three minutes preparation time, so you can take notes, some paper and pen, but ….. I want you to take a topic first. Choose a card. ……. And tell us the number please.
St. This number?
In. Yes.
St. B2.2.
In. B2.2. So, paper, pencil and three minutes for preparation (E12)

In the first instance, the interviewer has decided to remind the student of the length of the preparation time, and in the second instance, it is the provision with pen and paper that is reiterated. The first reiteration probably serves the purpose of emphasis, ascertaining that the candidate has heard the information. On the second occasion, it is probably the distance between the first mention of pen and paper and the actual need to use them that prompts the reiteration. Those two examples clearly seem to aim at clarifying the preparation phase procedure. All in all, 7 interviewers reiterated information in the lead-in.

Other types of additional elements include:

1. promise to let the candidate know when the preparation time is up;

E.g. Ok, starting from now you have three minutes to prepare, you can take notes, here is a pen and I will tell you when the time is up (E 4).

E.g. And now you have three minutes. And I will let you know when three minutes has passed (E20).

Nine (2/7) interviewers added that element to the lead-in. This kind of comment seems to add to the candidate's comfort during the preparation time, as the remark removes the obligation of keeping time from the candidate, so he/she can concentrate on the topic. It may be worth considering making this an official part of the script.

2. advice not to hurry;

E.g. And starting from now you have three minutes, take your time and then you can talk. (E5).

This comment only occurred on two occasions, and although it seems to be aimed at reducing the candidate's anxiety, is actually misleading, as the candidate is not at liberty to utilise unlimited time and is bound by the 3-minute margin.

3. questions regarding readiness to start;

E.g. Here is paper and a pencil to you and are you ready? (E2)

The above element occurred in but three interviews, but in each case, it seemed unnecessary, as, at that point (the candidate had taken the topic, pen and paper and had comprehended the task that had been given to him), the candidate was supposed to start preparing for the monologue. To enquire whether he was ready could only be misconstrued as a request to start speaking.

4. use of the candidate's first name;

E.g. In. Right aaa Zenja, now I'd like you to speak on a topic for two minutes, before you talk you have three minutes to think about what you are going to say (R6).

E.g. In. Good. Vika, now I would like you to speak on a topic for two minutes. Before you talk you have three minutes to think about what you are going to say. You can make some notes if you wish. Do you understand? (R7)

Candidate name occurred in 8 interview lead-ins (all recorded in Russian schools). When questioned in private conversation with interviewers (not recorded, and can therefore only count as anecdotal evidence) why they included the candidate's first name (and often not just the first name but the diminutive form of the first name) in the instructions, the interviewers affirmed that this was done in order to establish rapport with the candidate, reduce the formality level. It is interesting to note, that it seems to be culture-specific and gender-specific, as this feature of the interviewer's language manifested itself only in the interviews conducted in Russian schools and all the interviewers were female.

5. stating the name of the topic area;

E.g. In. Well, I see, now please, choose .. your … task card. So, A6.1.Hobbies and culture. You have 3 minutes in order to prepare for your monologue over there (R4)

There were 4 interviews where the topic was announced (all with Russian interviewers). Knowing the topic number is important for the interviewer and the assessor

during the examination, stating the general topic to the candidate does nothing but confuse him/her, so there is no need to include that.

6. giving the candidate the option to start when he/she is ready;

E.g. You have 3 minutes in order to prepare for your monologue over there. You've got some paper and a pen in order to prepare yourself. When you are ready, you may start with your monologue. (R4)

Only two interviewers (1/1) were observed to include this information in the lead-in. The candidates at the national examination speaking test have the option to forgo the preparation phase altogether and start speaking immediately, or not to use all the three minutes if they are so inclined. However, while designing the lead-in, it was decided not to include that element in it to encourage the candidates to use all the allotted time not to run out of ideas during the monologue presentation. From this point of view, the above comment is unnecessary.

7. misinformation;

E.g. Now I would like you to speak on a topic for at least two minutes. (E13)

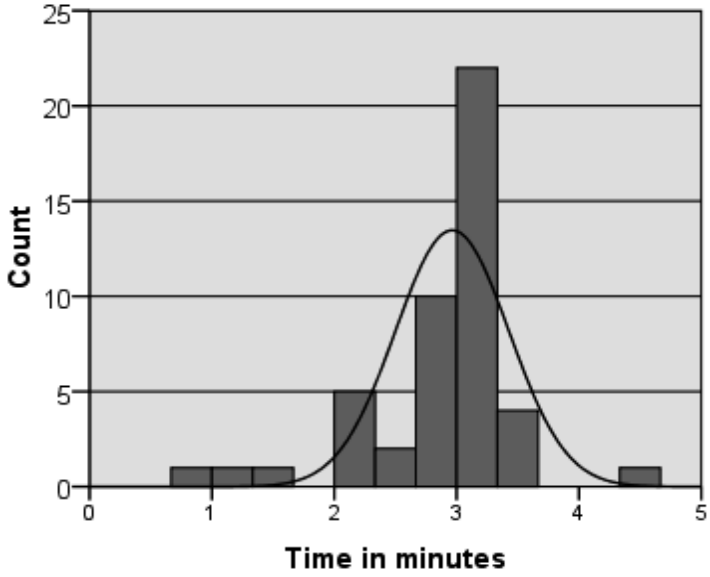
Including the above remark in the instructions, the interviewer inadvertently gives the candidate false information about the expected length of the monologue – two minutes minimum – whereas the scrip requires just two minutes (and not longer). By expecting the candidate to speak longer she will place the candidate in an unequal position compared to other candidates.

The above discussion reveals quite a variety in terms of the amount and the kind of information the interviewers give the candidate in as small a section of the oral interview as the lead-in.

4. 6. MONOLOGUE PREPARATION TIME

According to the lead-in of Task1, the candidates had the opportunity to spend three minutes preparing for their up-coming 2-minute monologue. The candidates also had the option to forgo the opportunity if they felt they were ready to commence without preparation. The interviews investigated were analysed for the actual time candidates spent preparing for the monologue. Here, too, for reasons discussed above (cf. Section 4.2.), 47 of the 50 interviews could be included in the analysis. The shortest time taken was 41 seconds (R22) and the longest time allowed was 4 minutes and 24 seconds (E19), the average time spent on the monologue preparation being 2 minutes and 51 seconds. The results have been summarised in figure 2 below.

Figure 2. Monologue preparation time.



The average time falls within the time envisaged, but over a half of the candidates had the opportunity to use more than 3 minutes for preparation. This may have stemmed from the ambiguity in establishing the start of monologue preparation time discussed above, or from the general leniency of the interviewer with regard to the rigour of timekeeping.

4. 7. INTERVIEWER LANGUAGE DURING TRANSITION FROM PREPARATION TO MONOLOGUE

Transition from monologue preparation phase to delivering the monologue could be either interviewer or candidate initiated, i.e. when the three minutes were up, the interviewer could stop the candidate and instruct him/her to start speaking. On the other hand, the candidate could stop preparation at any point during the 3-minute phase himself/herself. Of the 46 interviews analysed, 40 represented cases where the interviewer stopped the preparation. There was one case in Estonian schools and 5 cases in Russian schools where the transition was candidate initiated.

The script provided the interviewer with the following transition sequence:

> Alright, remember you have two minutes for speaking. I'll tell you when the time is up. Please start speaking now.

The sequence had the following functions:

1. to stop monologue preparation (Alright),
2. to remind the candidate of the monologue length (remember you have two minutes for speaking),
3. to assure him/her that the interviewer will keep time (I'll tell you when the time is up),
4. to request to start the monologue (Please start speaking now.).

As above, an attempt was made to give the candidate as much information as possible and be clear about the expectations both to the candidate and the interviewer. The last sentence, in addition to giving a clear signal to the candidate to start speaking, also served as a signal to the interviewer to start the monologue timer.

All the interviews were analysed for the elements listed above, in addition to which a study was conducted regarding the types of changes introduced by the interviewers. The results of the investigation are summarised in the table below.

Table 10. Element representation in the monologue.

No.	Element	Present in Transition	Absent in Transition
1.	stop monologue preparation	18/16	3/9
2.	note on monologue length	10/17	11/8
3.	time-keeping	8/15	13/10
4.	start the monologue	18/19	3/6
5.	order changed	1/0	20/25
6.	wording changed	12/5	9/20
7.	additional information	9/12	12/13

All in all, there were just 13(8/5) interviewers, 28.3% of the total, who followed the transition verbatim without any changes, the rest – 71.7% – conducted the transition with alterations. Studying the recordings of the interviews, the most laconic signal to the candidate just consisted of either a direct or an implied request to start:

(1) E.g. In. Well, could you start please. (E2)

(2) E.g. In. Ok welcome. (R18)

In the first instance, the interviewer asks the candidate to start speaking and in the second, the interviewer resorts to what seems to be an elliptical form of the full phrase 'you are welcome to start'. By reducing the transition to such a short phrase, the interviewers must have been guided by the consideration that there was no further comment necessary because the candidates had been advised about the interview procedure, so repeating it must have seemed superfluous. It is imperative however that all candidates be given the same information, irrespective of how tedious or

repetitive the procedure may seem to the interviewer. Only then can comparisons be made between candidate performances.

As was pointed out above, the transition could have been either interviewer or candidate initiated. When the transition was candidate-initiated (out of 46 interviews there were 6 such instances), an interesting phenomenon manifested itself. On no occasion did the interviewer tell the candidate about the time allowed for speaking or time keeping, neither did they mark the time when the monologue was to begin. Although such instances had been discussed during training, it was interesting to note that the interviewers left the candidate to their own procedural devices if the latter took the initiative.

In the overwhelming number of cases – 34 (18/17) – the transition was interviewer – initiated and conducted, resorting to the wording suggested in the script, i.e. using the marker 'alright'. Besides the six candidate – initiated transitions, there were 6 cases where the interviewer employed a different wording to initiate the transition. The nature of the changes will be discussed in the respective section below.

While putting together the transition sequence for the interviewers it was considered necessary to remind the candidates of the expected length of their monologue. Of the 46 interviews analysed, there were 27 (10/17) where the interviewer did that, and 19 (11/8) did not mention it, reducing the clarity of the procedure. Here, the Estonian interviewers seemed to be almost equally divided whereas the ones who retained the comment in the transition among the Russian interviewers (68%) outnumbered those who did not (32 %).

Assuming the responsibility to inform the student of when the required time had elapsed divided the interviewers into two equal groups, there were 23 (8/15) interviewers who mentioned it and 23 (13/10) who did not. If the interviewer explicitly takes the responsibility of timekeeping, he/she will hopefully again deflect some of the examination-related anxiety, as the candidate can then rely on the cooperation on the part of the interviewer to monitoring the length of his/her monologue.

The request to start the monologue was explicitly expressed in 37 (18/19) interviews, being the most frequently present element of all the listed elements above. Leaving out the 6 cases where the transition was candidate initiated gives us 3 more cases where the request was implied rather than explicitly stated:

(1) E.g. In. Well your time is up. (E3)

(2) E.g. In. So, now you have two minutes for speaking. (E19)

(3) E.g. In. (Name), … Remember, you have two minutes for speaking. (R6)

In the first case, the only signal the interviewer gives the student to start the monologue is the remark that the preparation time is up. This seems to serve as an implicit order/request to start his monologue. In both the second and the third instance, the interviewers, after the preparation time has elapsed, announce to the candidate how much time he/she has for speaking without any additional comment and this seems to be interpreted by the candidate as a starting signal. On no occasion do the students get

all the information they are entitled to by the script. What is more, the first interviewer, being the most laconic, comes across as autocratic and completely non-accommodating. One wonders if the amount of information and accommodation provided by the interviewer during the interview does have a bearing on the amount and quality of the candidate's presentation during the interview. A very preliminary correlation that could be observed in the current study (which was not set up to study the relationship between the interviewer language and the amount of candidate language produced during the monologue) was that the transition exemplified by (1) above resulted in the second shortest (50 seconds instead of two minutes) monologue in the study. This problem needs further study though.

Changes in the Transition

As was pointed out above, 71.7 per cent of the interviewers made some sort of a change during the transition. While investigating the changes, there were two main foci: first, to see if there was a change in the sequence of the elements suggested for the transition and if so, which elements changed their places. The second focus was on the wording of the script elements.

a) Changes in the Sequence

Comparing and contrasting the abovementioned elements in the script with those found in the sequences which the interviewers actually used, it could be seen that changes could only be found in the wording of transition elements. Only on one occasion was there an alteration made to the order of the elements:

E.g. In: So, you may start. You have two minutes. (E20)

Here, the interviewer, besides leaving out the bulk of the instructions the candidate is entitled to, first tells the candidate to begin and then alerts him/her to the time available. This order was reversed in the script for logical considerations. It was thought necessary to give the candidates all the information about the forthcoming monologue before actually asking him/her to concentrate on speaking.

b) Changes in the Wording

Studying the changes in the wording, several patterns emerge. The first, fairly widespread change involved the phrase the interviewers used to stop the candidate. Instead of using the recommended 'Alright', frequently found in other similar examinations, the interviewers prefaced it by telling the candidate that his/her time was up.

E.g. In. Name, the time is up. Mhmh Alright. ……. Remember you have two minutes for speaking, I'll tell you when the time is up. Please start speaking now. (R9)

Sometimes, the recommended phrase ('alright') was replaced by the phrase ('your time is up').

E.g. Your time is up, so you should start. (E1)

E.g. In. Well …. the time is up. Remember you have two minutes for speaking. Mhmh. (R14)

Of the 40 interviewer-initiated transitions, 18 (8/10) fell into this category phrase (45 per cent). It seems to indicate that those non-native interviewers may have found it hard to consider the given 'alright' a sufficient signal for the students and needed a more direct marker.

Another noticeable change was in the wording of the request to start. The script used a polite on record request 'Please start speaking now'. The interviewers, however, were found to change the modality at times:

(1) E.g. In. Alright. You may start. (E5)

(2) E.g. In. Your time is up, so you should start. (E1)

(3) E.g. In. Yes, shall we start. (E18)

(4) E.g. In. Your time is up now. Could you start speaking. (E21)

The first instance is permission rather than a request where the power balance has been tipped to the advantage of the interviewer, the latter granting the right to speak. The second instance is advice given with a moderate amount of urgency (Celce-Mursia 1999:144), and the interviewer's power position is no longer as strongly pronounced as in the first instance. The interviewer relies on an outside demand (the time being up) while giving advice, but can still be perceived as a person who has the authority to be in the position to give advice. The third instance seems to be an invitation and here, the interviewer appears to be on a par with the candidate, resorting to an inclusive pronoun 'we'. The fourth instance is a request but the interviewer has increased the politeness level to the point where the beginning of the monologue seems to have been left to the discression of the candidate, and as a result the power relationship seems to have been shifted so as to give a slight advantage to the candidate. All in all 9 interviewers of the 46 analysed (19.6%) made changes in the modality of the request.

c) Additions to the Transition

There were two other types of changes that could probably be classified as additions to the transition script, although these additions mostly also resulted in the change in the wording. The first change continues the trend manifested and discussed in connection with the lead-in above. This is the tendency, with the Russian female interviewers to add the candidate's first name in the script.

E.g. In. Julia, the time is up. Mhmh. Alright. ……. Remember you have two minutes for speaking, I'll tell you when the time is up. Please start speaking now. (R9)

In the above example, we can see that the interviewer is fairly faithful to the script, except for the comment about time and the inclusion of the candidate's name. Of all

the 25 interviews there were 6 instances (24 per cent) of the first name being introduced in the transition. The reasons can be assumed to coincide with the ones discussed above (cf. Section 4.4.).

The other addition to the scrip involves the enquiry about whether the student is ready to start the monologue or not. Occasionally, the latter is worded as a condition to start the monologue.

(1) E.g. In. Well, three minutes are over, are you ready?

St. Yes.

In. Remember, you have two minutes for speaking. I'll tell you when the time is up. Let's start. (R7)

(2) E.g. In. So, your time is up. Please, if you are ready you may start with your monologue. (R3)

Four interviews contain the above-mentioned query. It is unnecessary in the transition as it may give the candidate a false impression that he/she is only expected to start the monologue if he/she does not need any more preparation time. Given that this is not the case, the candidates are being misinformed. As it happens, none of the candidates responded to the question in the negative, thus they interpreted it as an interruption device/ a signal to start speaking, rather than an explicit concession. One wonders what the interviewer behaviour would have been had the candidates actually denied being ready.

All in all, it could be seen that all the interviewers recognised the need of a transition to facilitate a smooth interview flow. The need to be faithful to the scrip in this section is for the most part disregarded by the interviewers.

4. 8. MONOLOGUE MANAGEMENT

Once the candidates started their monologue, it was the interviewers' task to monitor the monologue time and to ask the follow-up questions provided by the script. The expected monologue length was 2 minutes and the requirement was for the candidate not to exceed this time limit. Analysing the interviews from the point of view of monologue length, interviewer behaviour could be viewed regarding their rigour of timekeeping. The plausible scenarios for handling the monologue on the students' part was either to keep going for the given length of time and be stopped by the interviewer once the time-limit had been reached, or, if the topic proved to be too challenging, stop speaking once the ideas had been exhausted somewhat before the time-limit had been reached. The figure and tables below illustrates the time spent on the monologue by the candidates involved in the 49 interviews.

Figure 3. Monologue duration.

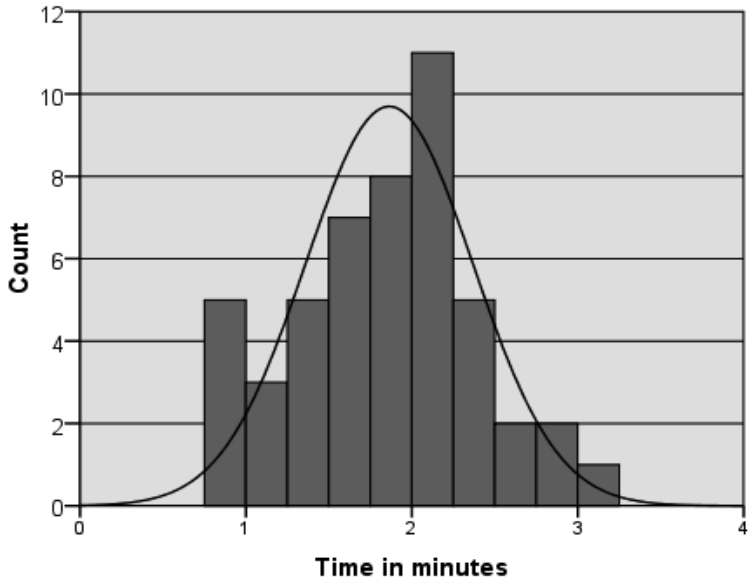


Table 11. Monologue time – Estonian schools.

	N	Minimum	Maximum	Mean	Std. Deviation
Monologue time	24	0 min 50s	2 min 48s	1min 56 s	0min 33s
Valid N (listwise)	24				

Table 12. Monologue time – Russian schools.

	N	Minimum	Maximum	Mean	Std. Deviation
Monologue time	25	0 min 45 s	3 min 14s	1min 43 s	0min 33s
Valid N (listwise)	25				

Comparison of the mean time spent on the monologue in both school types (tables 11 and 12) shows that the time spent on the monologue in the Estonian schools was generally longer than in Russian schools. Comparison of respective standard deviations reveals no significant differences between the two school-types, which was expected.

Another way to look at the statistics of the interview duration considering the school-type and the gender of the interviewer is reflected in the table below.

Table 13. Longest and shortest monologue times.

No.	Criterion	Longest	Shortest
1.	Overall monologue time (OMT)	3 min. 14 sec.	45 sec.
2.	OMT in Estonian schools	2 min. 48 sec.	50 sec.
3.	OMT in Russian schools	3 min.14 sec.	45 sec.
4.	OMT with female interviewers	3 min.14 sec.	50 sec.
5.	OMT with male interviewers	2 min. 48 sec.	57 sec.

The monologue was concluded by the candidate 24 times out of 49 (49%), whereas the Estonian students were more apt to end their monologues themselves (14) and Russian students were more often stopped by the interviewer (14) – 58.3 and 56 per cent respectively.

The number of interviewers who managed to keep the interview within the 2-minute time-frame was 29, just 59 per cent, 11 in Estonian and 19 in Russian schools. This means that of all the 49 interviewers, 20 allowed their students to exceed the time limit, in some cases by more than a minute (cf. Fig. 3 above). This could have affected the task given to the students in two ways: on the one hand, it could have made the task easier for some students as they had unlimited opportunity to demonstrate their language proficiency; on the other hand, it may have added to the task difficulty, as instead of two minutes they now were required to keep talking for a longer period of time since there was no interviewer signal to stop their monologue when the two minutes had expired.

If the candidate spoke for a shorter time than he/she had time for, and had not clearly indicated that he/she had finished his/her monologue (by saying, for example, 'that's all'), the interviewers had the obligation to verify the end by asking 'Is this all you wanted to say?' It was, thus, an optional question, only to be asked if the answer would be provided within the given 2 minutes. The interviewer behaviour differed in many respects here. Of all the interviewers, 17 asked the above question (10/7). In some cases, the question was justified and sometimes it was not. There were 5 instances where the student monologue length warranted the question. There were two occasions (E16, R5) when after a noticeable pause, when the interviewer asked the above question, the candidate proceeded with the monologue (E16 for 25 seconds, R5 for 44 seconds) providing a much fuller sample and still falling within the 2-minute time-limit. It is interesting to note, however, that, on some occasions, the question was asked even though the student had come to the temporal limit of his/her monologue. It was estimated that if 1 minute and 50 seconds of the monologue time had been spent, the question was not recommended any longer as the answer would have put the monologue over time. There were 6 such instances (E1, E7, E19, E23, R17, R22) where the question was not needed but was still asked, including an instance where

the candidate had already been speaking for 3 minutes and 14 seconds. On all these occasions, the monologue length went beyond 2 minutes. On the other hand, there were also 6 instances (E4, R3, R4, R9, R19, R23) where the question should have been asked because the monologue time fell below the 1 minute 50 seconds. The reluctance to ask the question may be related to the generally lower proficiency level of the above candidates, so the interviewers may have forgone the question in order to avoid the student giving a poorer sample than he/ she had already given, so it was a strategy the interviewers could (and did) use to save candidate's face.

After the candidate had finished the monologue, the interviewer's script prompted him/her to ask the candidate 4 follow-up questions. Asking the questions was obligatory and they were to be asked using the wording in the script, in the order they were given there. The reason for this was that the questions were graded progressing from easier to more difficult, each succeeding question requiring more sophisticated language (vocabulary and grammar) to answer. The interviewer could only skip a question if the candidate had already tackled the problem in his/her monologue.

The questions and answers were analysed in order to discover to what extent the interviewers followed the script, if and what kind of alterations were made and if any peculiarities could be detected in the interviewer behaviour there.

The overwhelming majority of the interviewers (46 out of 50) used the questions in the order they were given using the script wording. There were 4 interviewers (E4, E5, R8, R12) who did not limit themselves to the script questions but were rather guided by the candidate answers. There were two patterns that emerged here: some interviewers seemed to get carried away by what the student had previously said and wanted to know more (e.g. instead of 4 questions E5 asked 8 questions). In all those cases (E4, E5, R8) additional questions added to the task difficulty, as more time was spent on the questions for no obvious reason other than interviewer curiosity. On one occasion (R12), the additional questions seemed to stem from the interviewer's attempt to boost what appeared to be a fairly poor candidate performance. Rather than give the candidate sufficient time to respond to the question and if no answer was provided, move on to the next question, the interviewer resorted to paraphrased questions, attempting to make the task manageable to the candidate.

E.g. (1) In. mhmh. Thank you. Now, (name), I would like to ask you some questions. Have you ever saved money for something? (**Script question 1**)
(2) St. I have (long silence) I work.
(3) In. What kind of things have you ever saved money? (Not a listed question). For bicycle or ….for your dog, toys?
(4) St. Toys
(5) In. What kind of toys? Dolls or puzzles?
(6) St. Puzzles.
(7) In. Puzzles. Mhmh, (name), what do you think of young children having their own bank accounts? **(Script question 2**)
(8) St. (long silence) I don't know.

(9) In. Is it better for them?
(10) St. Better.
(11) In. Mhmh, (name), when do you think young people should get their first jobs? (extremely slow speech). (**Script question 3**)
(12) St. (long silence) I think our town.
(13) In. What kind of jobs?
(14) St. (long silence) Work in shop, work in garden (silence)
(15) In. (name), in Estonia, are the best paid people the ones who work hardest? (**Script question 4**)
(16) St. best.
(17) In. Do you agree with it?
(18) St. Yes I do.
(19) In. Can you explain?
(20) St. Not.
(21) In. Thank you.

In the example above, the candidate clearly struggles with the question comprehension and the interviewer resorts to a multitude of strategies to facilitate understanding and rapport with the candidate: instead of asking just the four script questions, she introduces additional questions (line 3, 9, 13, 17), makes them short and simple (line 5, 9, 13, 17) suggests answers (line 3, 5), uses slow speech (line 11), gives the candidate ample time to respond to the questions (line 2, 8, 12), uses the candidate's first name (line 1, 7, 11, 15) and backchannels frequently (lines 1, 7, 11).

Backchannelling is vocal indication to the partner that one is listening while the other is having a longer turn (Yule 2000:75). Vocal signals in the form of 'mhmh' were frequently used by the interviewers to provide that feedback. The signal sometimes seemed to be utilised as confirmation of the correctness of the answer, however: when the interviewer seemed to be especially pleased with the candidate's response for one reason or another (content, grammaticality perhaps), the signal was used with a particular approving intonation. Although interviewers had been instructed not to offer any comment during the candidate performance either explicitly or implicitly, the above reaction could at times definitely be construed as praise (there are 7 instances of such interviews).

Another noticeable pattern emerged when students, having failed to understand the question when the interviewer first asked it, requested that it be repeated. While it is reasonable to assume that the interviewers would repeat the questions using a somewhat slower tempo and a clearer articulation (Lazarton 2002:128), the interviewers involved in the current study also sometimes resorted to extremely slow speech unnaturally articulated, so that virtually every word stood on its own. There were 7 instances in the question and answer section where the candidates asked the question to be repeated, in 6 of them (E10, E11, E14, R2, R18, R23) a slightly slower tempo was observed whereas 1 displayed unnaturally slow speech (R21). Related to the above is the tendency to resort to slow speech in general while asking questions, compared to for example to the speed of speech while giving instructions for the up-coming tasks.

4. 9. INTERVIEWER LANGUAGE DURING THE LEAD-IN TO TASK 2

The management of the second part of the interview required completion of multiple tasks on the part of the interviewer:

1. lead the candidate into task 2 using the script provided,
2. keep time while the candidate was preparing for the role-play,
3. assume the role specified on the interviewer's cue card and respond to the candidate's queries and comments,
4. close the interview once the candidate had finished the role-play.

On the one hand, the interviewer continued to have the role of the manager of the interview, securing adherence to all the procedural requirements (e.g. supply instructions, keep time). On the other hand, the interviewer also had to assume the role of the interview participant, changing his/her mode of communication depending on the role requirements. While the interviewer role-cards provided the information necessary for an adequate participation in the role-play (complete sentences with a fair amount of detail), the interviewers had been instructed during training to modify their responses depending on the specific candidate questions. Thus contrary to what they had been instructed to do managing the introduction, task 1 and the lead-in to task 2 (acting as interviewers), they were at liberty to improvise here (acting as role-play participants). That dichotomy was expected to pose a challenge for some interviewers. Current research was interested if indeed a problem would arise while making a distinction between the two roles.

But first, the interviewers' conduct during the task 2 lead-in was investigated. Once the candidate had responded to the interviewer's questions in task 1, the interviewer signalled the end of that task by the marker 'thank you'. The marker was used without fail by all the interviewers in both school-types. The role-play was to introduced implementing the following script:

> Let's move on to the next task.
> Here is a card with a task on it. Please read it to yourself. You have one minute to think about it. I'll tell you when the time is up.
> (After one minute has elapsed). Could you start the role-play now.

The script was put together so that it would be clear and logical for the candidate; for that reason, simple vocabulary and short sentences were used, as with other parts above. For the analysis of the interviews, a similar strategy was used as with the previous sections of the script, i.e. the interviewers' language was compared and contrasted with that of the script, element by element, to discover the degree of adherence and the nature of alterations. The lead-in can be seen as having the following functions, represented by respective elements:

1. signalling change of task (let's move on to the next task),
2. introducing task 2 (Here is a card with a task on it),

3. instructions regarding manner of preparation (Please read it to yourself),
4. informing the candidate of the preparation time (You have 1 minute to think about it),
5. assuring him/her that the interviewer will keep time (I'll tell you when the time is up),
6. requesting to start the role-play (Could you start the role-play?).

The results of the interview analysis can be summarised in the table below.

Table 14. Element representation in task 2 lead-in.

No.	***Element***	**Present in Task 2 Lead-In**	**Absent in Task 2 Lead-In**
1.	Signalling change of task	22/23	2/2
2.	Introducing task 2	22/25	2/0
3.	Manner of preparation	19/21	5/4
4.	Note on preparation time	24/25	0/0
5	Time-keeping	16/17	8/8
6.	Request to start the role-play	18/21	6/4
7.	Order changed	0/2	24/23
8.	Wording changed	13/11	11/14
9.	Additional information	13/11	11/14

A cursory look at the table above reveals a more adamant adherence to the given script than to the scripts of the previous parts. The required elements are more often present and they are more frequently delivered to the candidates using the script wording than we saw with previous sections of the script. We can see, for example, that all the interviewers inform the candidate of the preparation time available to them, and all but two interviewers inform the candidate of the upcoming task Relatively fewer interviewers (4) explicitly signal the transition to the next task, and the variance increases with regard to the number of interviewers who tell the candidate how to prepare for the role-play (9) and especially regarding those who do not mention time-keeping (16). The figure representing the number of those interviewers who ask the candidate to start the role-play may be misleading in that it was not always the interviewer who stopped the preparation. Indeed, of the 10 cases where the request to start the role-play was missing, there were 8 where the beginning of the role-play was candidate initiated (and thus no signal was needed), and only 2 (E4, E23) where the interviewer should have used the phrase.

Besides maintaining the procedural steps envisaged by the script, it is also noticeable that the interviewers presented the information in the order the script suggested, which also suggests that the candidates were exposed to similar/ comparable conditions during the speaking test. As can be seen from the above table, 47 interviewers maintained the given order and there were only 2 interviews where changes were detected. A closer adherence to the script can perhaps be explained by the fact that

role-play as a task has a long history in the English language national examination tradition and the interviewers were more familiar with it. Task familiarity may have afforded them more time to focus and consequently adhere to the procedural demands.

The fact that 47 interviewers kept to the suggested order does not mean however, that all those interviewers kept all the elements in their lead-in. Of the 49 interviewers, there were 23 (12/11) interviewers who had all the required elements in their lead-in. To these we may add the interviews where the role-play start was candidate-initiated (5) and the final phrase was thus superfluous. This leaves 22 interviews where certain elements were left out.

The shortest lead-in to the role-play was achieved with just 3 elements, compared to the 6 in the script.

E.g. In. Ok. Thank you. Let's move on to the next task. …. <u>You have a minute to look at it and then to ask me some questions.</u>
pause
In. Alright you may start. (E5)

In the example above, of the 6 elements listed in table 14, only 1, 4 and 6 have been utilised. There is an exophoric reference to the task-card ('you have a minute to look at it'), which becomes clear though the action of the interviewer giving and the candidate receiving the card, information about the available time and permission to start. The additional comment, 'to ask me some questions', effectively reduces the task from a role-play to a considerably simpler question and answer task. The interviewer also inadvertently misinforms the candidate by saying that the candidate has a minute to accomplish both familiarising herself with the task and asking the questions. So, in effect, not only is the lead-in devoid of necessary pieces of information, it is incorrect and confusing.

All in all, there were 11 interviews (7/4) where only three of the six elements were communicated. The only element that was never dropped was the amount of time available to the candidate, all the other elements were dropped with a varying degree of frequency, as was shown in the table above. The most frequently dropped element was the interviewer's responsibility to keep time (16). A similar tendency manifested itself in the transition from monologue preparation to monologue presentation. There, too, the respective remark often seemed unnecessary to the interviewers. This may show the interviewer's unawareness of how they help to construct discourse and how much what they bring to the interaction influences candidate behaviour. In this case, it is rather what the interviewers do not bring to the interaction that sets the tone of the interview. By being laconic and enigmatic in their instructions – withholding information from the candidate – the interviewers may be perceived in two different ways, both of which are detrimental to the candidate performance. On the one hand, the interviewer may be perceived as being imperious and autocratic, expecting the candidate to be aware of the procedure before coming to the exam room and consequently not needing any detailed instructions. This, as we saw above, may result in misunderstanding and second-guessing on the candidate's part. On the other hand,

laconic, careless instructions may signal to the candidate that the upcoming task is not very important and whatever the candidate says may be construed as sufficient. On both occasions, the candidate will not give of his/her best, on the first occasion, because he/she is confused and misinformed, and on the second occasion, because he/she is not trying hard enough.

Although a greater degree of adherence to the script can be noticed in this section, alterations could still be observed here just like in previous sections. And just as above, the changes usually took the form of

a) changes in the order of the information given,
b) changes in the wording of the instructions,
c) additional information given to what was specified in the script.

All these changes can be further discussed and illustrated.

a) Changes in the Role-Play Lead-In Order

Changes in the order of information given were very rare – just two instances. One of the instances, however, is interesting from the point of view of how, if the information is not clearly worded, and presented, it might lead to misunderstanding and unexpected behaviour on the part of the candidate.

E.g. In. Mhmh. Thank you (name). And this is your role-play card. You have one minute to read the task on it.
St. (starts reading out the role-play card aloud)
In. You can read to yourself.
St. Ah. (Chuckles).
pause
In. Well, could you start the role-play now. (R6)

According to the script, the interviewer was to first ask the candidate to read the role-card to him/herself and then spend time thinking about it. Unhappily for the student, the interviewer first mentions the time allowed for preparation and follows it up by the need to read the task, never noting the need to read the task to one-self or the advice to think about it. Consequently, the candidate assumes he is expected to read out the information on the card. This leads to an additional comment (you can read to yourself) from the interviewer, which should have been in the instructions in the first place, and a slight embarrassment on the part of the candidate.

b) Changes in the Role-Play Lead-In Wording

One of the most frequent changes that could be observed was in terminology, in what the script called the 'card with a task on it'. The substitute phrases used were 'role-card' (e.g. R12, R13), 'cue card' (e.g. E1, E2), but also 'role-play card' (e.g. R7) and, on one occasion, an interesting 'play-card' (R11). In the instances where the alternative term was used, the part in the script mentioning the task was dropped.

E. g. In. Thank you. Let's move on to the next task. Here is your cue card and … you have to read it to yourself and you have one minute to think about it and I'll tell you when the time is up. And no note-taking.
Pause.
In. Could you start the role-play now? (E3)

E. g. In. (Name), thank you. Let's on aaaaa, let's move on to the next next task. Here is card, a role-play card. Please read it to yourself and think about. You have one minute for preparation of interview.
Pause.
In. Mhmh. Could you start the role-play now. (R10)

It is noteworthy that 18(6/12) interviewers consider it necessary to mention the name of the task – role-play – at the very beginning of the lead-in although the script does not mention it explicitly until later in the instructions. It seems to add clarity to the instructions. For that reason, it may be worthwhile editing the script by including it in the lead-in from the very start. The reworded first sentence of the role-play lead- in could sound as follows: Let's move on to the next task, the role-play.

A further frequent change is in the formulation of the request to start. As we saw in the part where the interviewer prompts the candidate to start the monologue (cf. Section 4.7.), some interviewers change the overall tone of the interview by wording the invitation to start the role-play not as a request but rather as permission to start speaking,

(1) E.g. In. So, thank you. Let's move on to the next task. You'll be given a card in order to prepare yourself for one minute for your role-play Hobbies and culture A6. Now you've got one minute on order to prepare the role-play. When you're ready you may start with the role-play.
Pause
In. It's ok. You may start. (R3)

(2) E.g. In. Thank you. Let's move on to the next task. …. Please … take…. your role play card. ……. Here it is. You have one minute to prepare. I'll tell you when the time is up.
Pause.
St. I'm ready.
In. Ok, so you can start … the role-play. (R22)

In the first example, the interviewer stops the candidate's preparation and prompts him to start the role-play by granting permission. The meaning of the sentence, 'it's ok', is not pragmatically clear, as the candidate does not seem to require any kind of reassurance (for which the phrase is usually employed). The formality level of the lead-in is further increased by employing passive constructions instead of the more neutral active ones. In the second example, the permission is a response to the candidate's wish to start. In both cases, the power scale is tipped slightly towards the interviewer.

Other types of wording are used to invite the candidate to start the role-play which are more inclusive than the examples discussed above, where the interviewer's wording implies a shared responsibility for the completion of the role-play.

(1)E. g. In. Thank you. Now we move on to the next task. And that is the role-play. So here is a card with a task on it, please read it to yourself you have one minute to think about it. And then you start the role-play.
Pause
In. Well, let's start. (E19)

(2) E.g. In. Mhmh. Thank you. Let's .move .on .with our .task. Give it to me. And ….. so … aaaa. Here is the task and your role play. Please read it to your self. You have only one minute to think about it. I will tell when the time is up. Ok.
Pause
In: Ok, (name), we can start. (R17)

Both examples above illustrate cases where the interviewer appears to consider himself/ herself a task participant, a partner to the candidate, rather than an interviewer. Using the inclusive 'let's' (1) and 'we' (2) seems to aim at establishing companionship, a more stress-free atmosphere. The phrases do not explicitly state, however, who has to take the initiative and start the role-play, which is why a more direct wording, suggested in the script, is to be preferred.

In this section, too, interviewers were sometimes noticed to prefer synonyms to the phrases suggested in the script.

(1) E.g. In. Thank you. Now you have one minute to think about … your… task and I'll let you know when your time is up and then you start the role-play. (E20).

(2) E.g. In. So, thank you. Let's move on to the next task. You'll be given a card in order to prepare yourself for one minute for your role-play Hobbies and culture A6. Now you've got one minute on order to prepare the role-play. (R2)

In the examples above, the interviewers have opted for a synonym instead of the wording suggested in the script. In the first example, a phrasal verb has been used ('let you know' instead of 'tell'), and in the second example an altogether different verb has been chosen ('prepare' instead of 'think about'). Both lead to a change in register, but in the second case, there is also to a change in meaning. With the role-play, the candidate is literally only allowed to think about the given task, he/she cannot take notes of questions he/she might want to ask or plan statements or stories. Using the verb 'prepare' in two meanings in the lead-in is unnecessary at best and confusing at worst. The first instance ('prepare yourself') seems to refer to an unpleasant approaching event, and the second instance ('prepare the role-play') implies that the candidate will either act out a memorised text or would have to prepare a role-play task himself/ herself (i.e. task development).

The same can be observed with cases where the interviewers replaced the phrase 'think about it' by 'look at it'.

E.g. In. Ok. Thank you very much. Take a look at this card. You have a minute to look at it and then to ask me some questions. (E4)

E.g. In. Ok. Thank you. Let's move on to the next task. …. You have a minute to look at it and then to ask me some questions. (E5)

In both cases, the interviewers have limited the lead-in instructions to the minimum along with changing the essence of the task. The role-play is never mentioned, instead, the candidate is just expected to ask questions. But more than that the explicit instruction to 'think about' the role is replaced by vague directions to 'look at [the card]', which render the whole lead-in obscure and imprecise. The confusion described above would have been avoided, had the interviewers adhered to the script given.

c) Additions to the Lead-In Script

Some interview transcripts reveal additions that particular interviewers have made to the role-play lead-in script. The first addition concerns note-taking during the role-play preparation. There are 5 (2/3) interviews that have the respective remark.

(1) E.g. In. Thank you. Let's move on to the next task. Here is your cue card and … you have to read it to yourself and you have one minute to think about it and I'll tell you when the time is up. And no note-taking. (E2)

(2) E.g. In. Ok, thank you. Let's move on to the next task. I will give you a card with a task on it and I need to have your notes, note-taking is not allowed in this part. You have one minute to think about it. And I'll tell you when the time is up. (R25)

All the interviewers who have included this comment in the role-play lead-in script seem to find it necessary to emphasise the contrast between this part (task 2) and the previous part (task 1) of the interview with regard to note-taking. While the remark in the first example appears as a kind of an afterthought, then in example two, the contrast is made explicit ('not allowed in this part').

Another element that several interviewers seem to add to the role-play lead-in is an explicit remark on the candidate's obligation to start the role-play.

E.g. In. You can give it to me. Thank you. Now you have one minute to think about … your… task and I'll let you know when your time is up and then you start the role-play. (E20)

E.g. In. So, thank you. Let's move on to the next task. You'll be given a card in order to prepare yourself for one minute for your role-play Hobbies and culture A6. Now you've got one minute in order to prepare the role-play. When you're ready you may start with the role-play. (R3)

Of the 49 interviews analysed, 6 (5/1) interviewers added the respective element, the Estonian interviewers being mostly the ones to add it. It may have been motivated by the desire to add procedural clarity.

Other additions in this section resemble the ones that have already been discussed in connection with several preceding parts. These include, adding the topic number to the role-play lead in, and using the candidate's first name.

(1) E.g. In. Ok. Thank you. Let's go on to the next task. Task two. You have a card with a task on it. Your number is four two. You have a card with a task on it. Please read it to yourself and you have one minute to think about it. I'll tell you when your time is up. And no note-taking at this stage. You can't take notes. (R24)

(2) E.g. In. Maria, thank you. Let's on aaaaa, let's move on to the next next task. Here is card, a role-play card. Please read it to yourself and think about. You have one minute for preparation of interview. (R10)

Both features can only be found in the interviews recorded in Russian schools. The general topic theme was mentioned in five different interviews (R3, R4, R5, R23, R24) and the candidate's first name was found in 7 cases (R6, R7, R9, R10, R11, R13, R17). The reasons and commentary for both findings, discussed above, are valid for the cases found in this section of the interview as well. It should be reiterated that using the candidate's first name may be believed to foster a friendly atmosphere during the interview and may thus assist the candidate to give of his/her best. Adding the general topic theme to the role-play lead-in can only cause confusion as stated above and unnecessarily increases the interviewer speaking time.

The final addition to be discussed in connection with the role-play lead in is also something that appeared earlier, in the monologue preparation phrase. There are interviews where the interviewer seems to stop the preparation for the role play be resorting to the question 'Are you ready?'.

E.g. In. Ok. Thank you. Let's move on to the next task. And his is your role play card. Please …… take this and you will have …… one minute ….. one minute to prepare and I will tell you when your time is up.

Preparation time used: 00.55.

In. (stops preparation at the above time): Are you ready? (E23)

There are three interviews (E23, E24, E25) with the above feature. On all occasions, the interviewers stays within the allowed preparation time (1 minute) and stop the candidate resorting to the question above when the time has elapsed. The candidates in all cases seem to interpret this as a signal to stop rather than a genuine question requiring an answer. None of the candidates gives a negative answer to the query, they all answer in the affirmative and start the role-play immediately. From the point of view of clarity, and to avoid situations where the candidate might interpret it as an opportunity to get more time by responding to it in the negative the interviewers should have used the more straightforward language suggested by the script. Related to this is another approach, where the respective element was inserted in the earlier section of the role-play lead-in.

E.g. In. So, thank you. Let's move on to the next task. You'll be given a card in order to prepare yourself for one minute for your role-play Hobbies and culture A6. Now you've got one minute in order to prepare the role-play. When you're ready you may start with the role-play.
Preparation time used: 1.11
In. It's ok. You may start. (R3)

Here the interviewer gives conflicting instructions to the candidate. On the one hand, the one-minute time limit is announced in line 3 in the example above, but this is cancelled out by the statement immediately following where the candidate is told to start only when ready, implying that the student has control over the available time.

To sum up, similarly to the previous sections the alterations to the script are manifold. Unfortunately, the changes and additions made by the interviewer rarely add substance or clarity to the instructions. In most cases, their effect is the opposite, making the candidate second-guess what is expected of him/her during a particular section of the interview.

4. 10. ROLE-PLAY PREPARATION TIME

Role-play preparation time was set at one minute. The time was envisaged for the candidate to familiarise himself/herself with the context of the role-play and plan his/her own role. The time could not be exceeded but the candidate had the right to stop the preparation any time during that minute if he/she felt ready to begin the conversation. The actual time spent preparing for the role-play during the interviews analysed could be summarised as follows:

Table 15. Summary of the role-play preparation time.

No.	Criterion	Longest	Shortest
1.	Role-play preparation time (RPPT)	1 min. 21 sec.	28 sec.
2.	RPPT in Estonian schools	1 min. 10 sec.	28 sec.
3.	RPPT in Russian schools	1 min. 21 sec.	34 sec.
4.	RPPT with female interviewers	1 min. 18 sec.	28 sec.
5.	RPPT with male interviewers	1 min. 21 sec.	49 sec.

The time available to the given set of students varied considerably. The group can roughly be divided into instances where the interviewer stopped the preparation and the instances where the preparation was stopped by the candidate. The average time spent by the candidates on the role-play preparation was 69 seconds. Figure 4 below indicates that most students seemed to cluster around the time-slot of 50–60 seconds, which would be expected. It was interesting to note that boys spent generally longer on the preparation – .79 minutes, while the girls' preparation time average was .67 minutes.

Figure 4. Dialogue preparation time.

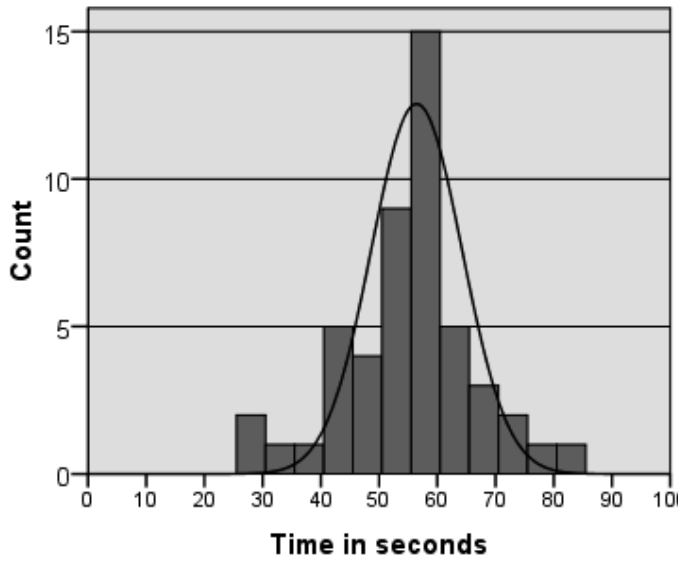


Table 16. Dialogue preparation time.

	N	Minimum	Maximum	Mean	Std. Deviation
Dialogue preparation time		28 sec	1 min 21 sec	55 sec	11 sec
Valid N (listwise)	49				

The number of interviews where the role-play preparation time remained within the appointed spectrum was 37 (17/20), 75.5 per cent, which means that in 24.5 per cent of the cases (7/5) the interviewer did not stop the preparation at an appropriate time and, consequently, the candidates had more preparation time compared to their peers.

As was pointed out above, the candidate could forego the preparation time allotted and start the role-play as soon as he/she felt ready. There were altogether 8 (4/4) instances when the candidate stopped the preparation. In all but three instances, the preparation time fell significantly below the 1-minute limit: on two occasions below 30 seconds (E11, E18) and on three instances below 45 seconds (E13, R14, R18). In the three remaining instances, the preparation time limit had either been reached (R19) or had already been exceeded (E4, E22), when the interviewer responded to the candidate comment that he/she was ready. The time-management during those three interviews was thus substandard as well.

4. 11. ROLE-PLAY MANAGEMENT

During the interviewer training, special emphasis was placed on the interviewer behaviour during the role-play. As stated above, the interviewers have a challenging task of acting as interviewers part of the time and assuming a role for part of the interview. All interviewers are supplied with an interviewer role-play cue cards that contain plausible responses to the candidate questions or comments. Predicting candidate remarks means second guessing up to a point, and this is why usually more information is given on the interviewer cue card than will be necessary in each instance of role-play enactment. Thus, the interviewers are expected to modify their answers and comments depending on the precise nature and tone of the candidate's question or comment, and not automatically read out all the information given for each question or comment on the cue card. In order to discover how the interviewers handle the information on their cue cards, the recorded role-plays were analysed for the naturalness of the interviewers' responses, the degree to which they did adapt their answers to the candidate's talk. Of the 49 interviews, 27 (15/12) represented cases where the interviewer made modifications to the cue card information and 44.9 per cent represented cases where the responses were read out exactly as the cue card stated. Characteristics of the interviewer responses have been summarised in the table below.

Table 17. Interviewer responses during the role-play.

School-type			Interviewer gender		Total
			Women	Men	
Russian	Read answers	Count	11	5	16
		% of Total	61.1%	71.4%	64,0%
	Adapt answers	Count	7	2	9
		% of Total	38.9%	28.5%	36,0%
	Total	Count	18	7	25
		% of Total	100,0%	100.0%	100.0%
Estonian	Read answers	Count	9	1	10
		% of Total	40.9%	50.0%	41,7%
	Adapt answers	Count	13	1	14
		% of Total	59.1%	50.0%	58,3%
	Total	Count	22	2	24
		% of Total	100%	100%	100.0%

It appears that interviewers in Russian schools read responses out unchanged more often than in Estonian schools – 64 and 41.7 per cent respectively. This finding is corroborated if the interviewer gender is considered as well. The proportion of female interviewers who adapted their responses was 59.1 per cent among the women in Estonian schools and 38.9 per cent of all the women who were interviewing in

Russian schools. The number of men in the study was too small to allow any generalisations. Thus it would appear that non-native interviewers in Estonian schools were either more informed of their task during the role-play and more knowledgeable of how to perform that task or they were more confident about their language ability to allow paraphrase and adaptation of the information presented in the interviewer cue cards. Conversely, interviewers in Russian schools seemed to need more support from the examination materials and relied on the cue cards verbatim more often than their Estonian counterparts.

Reading the answer verbatim seemed to be a strategy to use when the interviewer seemed to have a problem comprehending the candidate's question. Reading out the full answer in the cue card was probably employed in the hope that at least one part of it would be an appropriate response to the candidate's question. In extreme cases, the interviewer's answers seem to make up the bulk of the role-play, as in the example below.

E.g. In. Could you start the role-play now?

St. mhmh Hallo.

In. hallo.

St. What is your name?

In. My name is xxxxxx

St. Speak me please your history.

In. My history or this newspaper. History , newspaper history you mean, yah? It was launched as a free stapled colour newspaper in London in 1999.

St. What you popular places?

In. Popularity?

St. Yes.

In. Mhmh. It offers everything that a quality newspaper does, news interviews, features, TV listings, fun and games, sports, but for free.

St. …….. long silence. …….. How many best origins likes.

In. The concept comes from Sweden in 1995 Bill Anderson established metro international and started to publish the first metro newspaper in Stockholm.

St. Thank you very much.

In. Thank you. This is the end of the interview. (R12)

In the above-role-play, the interviewer clearly has trouble understanding the candidate because of the very low language proficiency level of the candidate. Except for the greeting, the very first question, and the subsequent request to speak about the interviewer's history, the candidate's questions are incomprehensible. For cases like this, a strategy should be developed where the interviewer would not proceed if the question or comment could not be understood and the role-play could be stopped when communication breaks down. As it happens, an impression of a complete interview is created although virtually no comprehensible communication happened.

A feature, which sometimes emerged in the interviewer behaviour during the role-play, was the interviewer's tendency to rush through the answers, read them at an unnaturally high speed, resorting to monotonous tone.

E.g. In. (stops the preparation at the above time) Could you start the role-play now?
St. Hello.
In. Hello.
St. How about some starters, what do you have for the starter.
In. You can have some salad from the salad bar.
St. Ok. But what about the main course. What is the main course?
In. You can have a nice steak which is really nice and tender, buuuuuuuut also you can also have beef chicken or fish.
St. And how about the sizes of portions?
In. There are small and big steaks.
St. Ok, aaa but what is for the dessert?
In. You should try our very best cheesecake.
St. Ok and what do you have for drink?
In. Juice and water.
St. How long will the service take?
In. The main course will take about 20–30 minutes.
St. Ok, and I'd like a small steak and for dessert I would have a cheesecake and water.
In. (no response)

In the above dialogue, the interviewer very quickly reads the answers from the cue card as they are written, resorting to the same falling intonation pattern with all the answers. One plausible explanation for the interviewer to choose an unnaturally quick speed for his/ her responses is to create the impression that he/ she is not holding the floor longer than the candidate, which would have countermanded the requirements set for the interviewer participation in the oral proficiency interview during the national examination. In this respect the interviewer is attempting adherence to the requirements. On the other hand, there is very little modification to the cue card information. The only time modification is attempted (see the underlined section in the above transcript), the interviewer very clearly gets disoriented (the adjective 'nice' is repeated in very close proximity, the vowel sound in the word 'but' is unnaturally elongated to hold the floor until she has had time to acquaint herself with the rest of the information on the card). After the turn described, no other alteration is made. Once the candidate has reached a decision there is no comment from the interviewer (there was none on the cue card), although an acknowledgement of some sort would be expected in respective situations in real life. This defeats the purpose of the second task where the candidate is expected to demonstrate his/ her communication skills. It is very difficult to estimate the exact number of cases within the given dataset, as we are analysing a single representation of a particular interviewer's interviewing style. Additional studies with numerous interviews from

the same interviewer should be analysed to put the claim on a more substantiated footing. There does seem to be preliminary evidence to suggest a behavioural pattern, as the variation of the speech speed and intonation within even one interview demonstrated notable variability from one part to another with some interviewers. In cases where the interlocutor is quite obviously disinterested and not committed to the task, the candidate will not be able to wholly commit either, as, rather than talking to the role-play partner, listening to his/her answers and responding to them, the candidate will just focus on getting the required questions asked, not paying attention to the responses he/she gets, i.e. not properly communicating. Such a behavioural interviewer pattern has an adverse effect on the candidate's performance and may affect his/her ultimate examination score.

4. 12. INTERVIEWER LANGUAGE DURING CLOSING THE INTERVIEW

The final section of the interviewer script required that the interviewer clearly indicate to both the candidate and the assessor that the end of the interview had been reached. Of the 49 interviews analysed, there are 3 (E2, E19, E20) that do not contain the respective section. To claim that the interviewers clearly did not signal the end of the interview in those cases would however be somewhat problematic as in each of those cases, the recording ends once Task 2 has been finished. Thus it is possible that the script was still followed although there is no recording of it. As this conflicts with the instructions the interviewers had, namely to switch off the recorder after the completion of the whole interview (also indicated in the script), the above-mentioned interviews are considered incomplete. Candidates need to get clear procedural information at the beginning, during and at the end of the interview. The information has to be the same for all the candidates. Failing to declare the interview finished deprives the candidate of the impression of completion, the feeling of closure.

4. 13. OTHER OBSERVATIONS

a) Interviewer Accommodation

One aspect that current research was interested in while studying interviewer behaviour during the national examination speaking tests was what strategies interviewers used in case candidates requested explanation of unfamiliar vocabulary they encountered on their task cards. An attempt had been made to compose the prompts so that the language on the task cards would be well within the candidates' expected language proficiency level. Thus, the candidates were expected to cope with the task without additional information. It was necessary, however, to see if the prompts did pose comprehension problems and if so how the interviewers handled it in case clarifications were sought. There were altogether 3 interviews where the candidate explicitly asked the interviewer to explain lexis (E11, E12, R12). On all occasions the interviewers provided a synonym as explanation:

E.g. In. Are there any household chores that your classmates avoid?
St. …… mmm can you ask again?
In. Are there any household chores that your classmates avoid (slows the tempo and pronounces every word separately).
St. What is chore?
In. Just some duties and tasks they have to fulfil.
St. Actually I don't know, I haven't asked them. Maybe if they are lazy. They don't they don't clean their room or something. (E11)

The fact that very few candidates seemed to require additional explanation may indicate that the exam design was successful while establishing the language difficulty level.

The above conclusion should be treated with care, however, because explicitly seeking clarification is just one of the strategies that the candidates could have resorted to while solving comprehension problems. This seems to be characteristic of candidates who are more confident about their own language skills. Candidates who are more insecure seem not to want to risk losing face by openly admitting (by asking the interviewer to explain) that they do not comprehend what has been asked. Instead, they resort to other strategies. An observable strategy is to respond with silence to questions that they do not understand. What the current research is interested in is to detect how interviewers reciprocate in such cases.

(1) E.g. In. Thank you, (name). Now I would like to ask you some questions. (Name), what is the <u>furthest destination</u> you have travelled to.
St. silence
In. Do you remember about your <u>longest travelling</u>?
St. Ah (understands now). My longest travelling was two years ago …. Saaremaa Island. Here … I was two days. This very beautiful, very magic place….
Where …. I see a lot of new … and I want to go … in this year, too. (R13)
(2) E.g. In. (Name), why do you think, there are <u>fewer female than male</u> in Parliament and politics. (mistakes on tape)
St. long silence.
In. <u>So many men but … women ….. less</u>
St. I think … the mens .,..they think …. fast, womans… they all decisions maybe they think a long time and when they (2 inaudible words) very well but sometimes it will perfect. But no one listen. Everyone speaking like this she is a woman and (1 inaudible word). (R9)

Leaving aside the interviewers' own language proficiency at the moment, we can see that, on both occasions, the candidate had trouble understanding particular lexis ('furthest destination' in example 1 and 'fewer females than males' in example 2). On neither occasion was the interviewer explicitly asked to explain the respective lexis, and by rights, should have moved on to the next question. Instead, the interviewers, interpreting the candidates' silence as non-comprehension, provide an explanation, thus resorting to an accommodating behaviour, a behaviour that makes

the task more accessible/ easier to the candidate. The accommodating behaviour leads to a response from the candidates – albeit fragmented and ungrammatical – that was not achieved with the original question. The level of interviewer accommodation is a feature of interviewer behaviour that should be agreed on within the framework of a particular examination to guarantee equal treatment of all candidates during testing. If the interviewer resorts to it with one candidate but does not do so with another, or if some interviewers utilise this while others do not, the candidates are ultimately in unequal conditions.

Another feature of the interviewer accommodation is varying the speech rate depending on the candidate proficiency. The feature was discussed in connection with instances when the candidate failed to understand the interviewer questions and requested repetition (cf. section 4.8.). It can, however, be observed in other sections as well.

E.g. In. Thank you.. Here is … your role-play .. card with .. a task on it.
Please. read it. to. yourself. You have one minute ….to think about it.
I'll tell you when the time.. is ..up. (R8)

The number of dots between the words in the given example indicates the approximate length of pause between them in the interviewer talk. Here the slowing of the tempo has happened during giving instructions for the role-play. There does not seem to be any obvious reason for slowing down the tempo other than wanting to make sure that the student understands properly. This strategy seemed to manifest itself more conspicuously in Russian speaking schools. Of the 25 interviews, 12 seemed to display occasions where the interviewer slowed down the tempo during the interview for the purpose of achieving clarity.

Interviewer accommodation could also be detected in cases where the candidate has failed to complete that task in the way the task card has prompted him/ her, and the interviewer, noticing that, tries to draw the candidate's attention to that. For example, at the end of task 2, the candidate is usually expected to come to some sort of decision with regard to the information he/she has obtained through the role-play. Failing to express that decision can be interpreted as task only partially completed by the assessor, and that, in its turn, may mean losing points on the scale of task completion. In some cases, the interviewers have been noticed to draw the candidates' attention to the missing part, thus trying to make sure the candidate completes the task, consequently leading them to achieve a higher rating.

E.g. In. Alright. Could you start the role-play now.
St. Hello.
In. Hello.
St. I heard you went to the language course in England. I want to go there, too. And I would like to ask you some questions.
In. Very good.
St. Could you tell me please where it actually took place?

In. Mhmh. It was in Oxford.
St. In Oxford, OK. And … how long did it last?
In. For two weeks.
St. Two weeks. ……. And what topics did you cover there?
In. British customs and traditions.
St. Ok What …. What about the accommodation?
In. We stayed with a family in Oxford.
St. Family.
In. Yes.
St. Ok. Mmmm aaa Did you take part in some ….. did you …. take part in some cultural activities.
In. yes, I did. There were trips to London and Stratford.
St. Ok and finally … I would like to ask you how much did it cost?
In. 20000 Estonian kroons.
St. Twenty thousand. It's not very few. Ok, thank you.
In. Good, (surprised tone) Is that all you wanted to say?
St. I think I would like to go there. I would collect some more money and I will go there.
In. Why?
St. I would like to study English more.
In. Alright, thank you. This is the end of the interview. (E9)

In the above interview, the interviewer is surprised to hear the candidate finish the role-play with, 'ok, thank you', although he/she has not expressly concluded the conversation with a decision, and expresses that with her intonation. She then prompts the student to continue with 'is that all you wanted to say?' which gets the necessary response from the candidate. All in all three interviewers were noted to resort to the above tactic (2/1).

b) Backchannelling

A typical feature of demonstrating to the partner that one is listening to what is being said is giving backchannel signals, which might be verbal or non-verbal. Typical non-verbal backchannels would be nods, smiles, gestures, etc. Verbal backchannels may take the forms of mhmh, uhuh, no, yeah, etc. Non-verbal backchannelling could only have been discussed with the current interviewers had the interviews been videotaped. The audio-tapes, however, did allow some conclusions regarding verbal backchannelling. Backchannelling was observed with the interviewers with varying degrees. Interestingly, the only forms of back channelling that was detected during the interviews in both schools was 'mhmh' and 'ok', as illustrated in the examples below.

E.g. In. How are you today.
St. I am fine.
In. That's nice to hear. Ok. Let's talk about …. Pets. Do you have a pet?

St. Yes I have two pets.
IN. Mhmh (approvingly).
St. I have one cat name Markko and a dog whose name is Isobel.
In. Mhmh. Why do people usually have pets.
St. I think aaaa because ……. Aaa pets are like friends and when people are lonely then pets make them feel I don't know … not lonely.
In. Ok, thank you. (E11)

E.g. In. (Name), do you agree that school is the student's second home.
ST. Yes, I agree because … aaa … students at school are six or seven … six or seven … mh … we at school aaa…so long time and at home … we just …eat and sleep and did home work.
In. mhmh. And how important is it to keep learning after starting a job?
ST …. ….. Can you repeat.
In. How important is it to keep learning after starting a job?
St. hm (children shouting in the background) … ….. ….. I think to start …. After education … better than work and learn together… cause aaa when you job oi when you work and learning you can't do something … maybe you can't
In. mhmh.(name), should schools teach students more theoretical or more practical skills?
St. I think more theoretical and maybe some small practical … because aaa school aaa help aa a future life.
In. Mhmh. Thank you. (R7)

All interviews were analysed for the occurrence of the backchannel signal 'mhmh', as this seemed to manifest itself more readily with some interviewers than with others. Table 18 below represent the findings concerning the respective backchannel signal.

Table 18. Backchannelling found in different gender interviews.

		Interviewer gender	
		Female	Male
Mhmh	Used	29	1
	% within Int. gen.	72,5%	10,0%
	Not used	11	9
	% within Int. gen.	27,5%	90,0%
Total		40	10
	% within Int. gen.	100,0%	100,0%

The statistical significance of the correlation was measured by chi-square-test. Correlation between bachanneling and the interviewer gender appeared to be statistically strong and significant ($p<0,000$, phi=0,51). The correlation between backchannelling and school type was not statistically significant (see appendix 4.1). The signal seems to have a dual function, though. It is mostly produced in monotone and, on such occasions, seems indeed to be utilised to maintain the flow of the interaction. Occasionally,

it seems to be the means of giving feedback to the student. In such cases, the signal is intoned with a slightly rising intonation, communicating approval of what has been said. It is overwhelmingly in Russian schools that the given function manifests itself – of the 13 interviews where the signal was found 7 interviews in Russian schools displayed the signal in the latter function (R8, R9, R10, R11, R12, R15, R17) as opposed to one interview among 16 found in Estonian schools (E11).

c) Correcting Mistakes

During the interviewer training, one of the requirements for the interviewers was that not only could they not correct candidate mistakes during the interview, they were not allowed to indicate through any means of body-language that a mistake/ an error of any sort had occurred in order not to prevent the candidate from displaying his/ her actual language proficiency level on the one hand, and not to increase their anxiety level, on the other. None of the interviews analysed displayed occasions where mistakes were corrected during the time the candidates were completing the tasks. However, instances of error correction were detected in the part where that candidate was asked to tell the interviewer and the assessor their topic number.

(1) E.g. In. Here is a pen and some paper. Please pick a topic. …. What is the number of your topic?

St. B three one.

In. <u>B three point (emphasised) one</u>. Now you have three minutes. I'll tell you when the time is up. (R2)

(2) E.g. In. Before you talk you have three minutes to think about what are you going to say … about the topic. Mhmh. You can make some notes if you wish. Here is a pencil and some paper. You can choose any card, please. Mhmh, mhmh. Olga, tell us, what's the number of your topic.

St. Two and two.

In. C (sii)

St. C two and two

In. point two.

St. C2.2.

In. nice, please. (R12)

In the first example, the interviewer corrects the mistake by modelling the correct answer, emphasising the missing part ('point'). In the second example, correction takes longer, the interviewer first correcting the letter and then the figures. The exercise is completed by the interviewer complimenting the candidate for getting the phrase right ('nice'). There were altogether 4 instances of interviewers correcting the students at this point of the interview (R2, R12, R13, R18), all in Russian speaking schools. This is a further indication of how challenging it is for the teachers to assume the interviewer's role, ignore the behavioural patterns which are acceptable for the teaching context but not during interviews.

c) General Level of Preparedness

It could sometimes be observed listening to the interview recordings that the interviewer seemed to be familiarising himself/herself with the interview materials while conducting the interview only. This manifested itself in long mid-sentence pauses while giving instructions for the task, getting lost in the script while asking follow-up questions, lengthy shuffling through papers while trying to find either monologue or role-play cards. For example, R3 proceeds to ask wrong follow-up questions which are not connected with the topic the candidate has been speaking about and then corrects himself by moving on to the correct set. R23 first allows the candidate to speak on the monologue topic relying on the prompt questions and then goes on to ask the candidate all the same questions that the candidate has just been speaking about and then proceeds with the follow-up questions, asking the student 9 follow-up questions instead of the 4 required. E22 forgets to ask one of the follow-up questions to task 1 and decides to still ask the question after the candidate has completed task 2. This was by no means a prevailing tendency (all in all 4 interviews stand out for features like this), but was all the more conspicuous for unprofessionality among the generally smooth flow of the interviews.

Interviewer preparedness also manifested itself in the ability to distinguish between different types of discourse necessary at different stages of the interview, This has already been discussed above in section 4.12.

An aspect of interview preparedness was knowing what to record if the candidate requested recording. As discussed at the beginning of the current chapter, there were 3 instances (E 23, E24, E25) where the interviewer only switched on the recorder at the time when the candidate started his/her monologue. As we have been discussing in the current thesis, the candidate response during the OPI can only be evaluated in the light of the part played by the interviewer in the interaction. Consequently, it is just as important to monitor the interviewer as it is to monitor the candidate to adequately evaluate the candidate performance.

d) Background Noise

During the interviewer questionnaire study, discussed in the previous chapter, questions were asked about the physical conditions of the national examination speaking test. There was a proportion of teachers who asserted that they had little say about where the interviews were conducted, which might mean poorer conditions for certain candidates compared to the others. While analysing the interviews above, background noise was one of the features that seemed to distinguish particular interviews from the rest. Of the 50 interviews analysed during this study, there were 10 (E1, E2, E10, E19, R7, R8, R12, R13, R17, R18), where some sort of a disturbing noise could be easily identified in the background.

Table 19. Background noise.

		Noise		Total
		Yes	NO	
School-type	Russian School	6	19	25
	Estonian School	4	17	21

The data reveal that there were marginally more Russian-speaking schools than Estonian-speaking schools where a background noise could clearly be identified to the point of being disruptive. This ranged from ticking noises (e.g.R17) to adults discussing something outside the exam room (e.g.E19) to children shouting during recess (e.g. E1). School management of the respective schools should be alerted to the need to provide quiet conditions for the national examination interview. Not only is it necessary to foster candidate concentration and consequently performance, but it may have a direct bearing on the assessment procedure. In the given cases, the noise level on some occasions was such that it could have potentially distracted the assessor, as it did the current researcher. Thus the assessor's attention could have been misdirected and consequently an uninformed judgement could have been made with regard to the candidate's performance. Moreover, should a second evaluation be called for, the aforementioned noise level could present a hazard for the assessment.

4. 14. CONCLUSION

The above analysis has allowed us to observe some of the general tendencies of interviewer behaviour during the national examination in the English language in Estonia and to illustrate those tendencies with concrete examples. The findings are valuable from the point of view of assessing the general level of national examination interviewer proficiency and point out the areas that need to be addressed during interviewer training and examination development.

It can be stated at this point of speaking examination development that there are interviewers who maintain procedural requirements without fail all through the interview, who manage to guide the candidates smoothly through the very stressful procedure while still obtaining all the necessary information. There were altogether 10 interviewers of the 50 analysed (8/2) who belonged to the above-mentioned group in the current study. This makes up just 20% of all the interviewers observed. If the NEQC database consisted of such interviews only, a huge step would have been taken towards achieving reliability of the speaking test, and, consequently, that of the whole national examination. Unfortunately, the number of those interviewers who deviated from the above group does not warrant that conclusion. Eighty per cent of the interviews reflected deviations from the standard, which means either complicating or simplifying the task for the candidate, making comparisons between candidates virtually impossible.

Analysing the interviewer behaviour from the point of adherence to the task they were set – to conduct oral proficiency interviews implementing the tasks given within the topic range envisaged by the national curriculum and the examination specifications, and to do that relying exclusively on the interviewer scripts – it appears that the first part of the task was completed commendably well. All but one of the investigated interviewers adhered to the overall oral proficiency interview structure: the candidates went through a three-phase interview, which included an introduction, task one (a monologue on a controversial statement, followed by questions), and task two (a role-play). On all occasions, the interviewers kept to the topics set for the respective tasks.

Adherence to the provided scripts was less successful and manifested a number of behavioural patterns:

- There is a general attempt to manage time during the interview, but the success rate varies from one interviewer to the next, with women generally being more successful than men as time-keepers.
- Interviews and their respective parts were generally completed quicker than pilot testing had suggested. This seems to have been achieved at the expense of shorter interviewer turns during interview management. The requirements for students in terms if adherence to the topic and the timeframe were much more rigorously observed than requirements set for the interviewers.
- There is a general tendency to preserve the script elements in the interviewer talk while managing the OPI, but the sequence and the wording of the elements display patterns of variation, resulting in the candidates having a varying amount of input during the speaking test.
- Script changes usually take the form of the change of script element sequence, omission of script elements, substitution of script element wording by another version or additions to the script wording.
- Interviewers resort to convergent accommodation techniques that aim at overall clarity of instructions to the candidate, such as reiteration of particular elements of the script, over – enunciation of instructions, questions or answers, slowing down the tempo, prompting task-completion.
- Interviewers use accommodation techniques to make the tasks more accessible to the students, such as simplification of questions by using paraphrase, asking additional questions about topics that students seem to be keen to talk about, finishing student turns.
- Interviewers vary in rapport-establishing techniques such as greeting, asking about the students' general welfare, backchannelling, etc., whereas some of the rapport-establishing techniques seem to be culture-specific (using the candidate's first name or the diminutive of the first name) or gender- specific (backchannelling).

- Interviewers use linguistic devices during the interview to signal varying levels of dominance. There are cases of both increasing and decreasing the formality level in the database.
- Interviewers modify the scripts to increase the level of directness. This is more visible with Estonian than with Russian interviewers.
- A generally low level of interviewer proficiency is manifested in the following behaviours: occasionally including misleading information in the script, in a fairly high level of script dependence (interacting with the script rather than the candidate), engaging in unnecessary behaviours (correcting mistakes) and the struggle to distinguish between the interviewer roles during the interview (managing the interview vs. asking questions vs. participating in a role-play), occasional substandard language use.
- Analysis of the results from the point of view of school-type seems to reveal a higher confidence level of both language and interviewing proficiency among the interviewers in Estonian schools and higher levels of insecurity in both areas among Russian school interviewers.
- Gender proportions in the current study were too imbalanced to allow many generalisations, but there seems to be a greater level of general adherence to the script demands on the part of female interviewers than of male. Female interviewers seem to be more oriented towards establishing rapport with the candidate than their male counterparts.

4. 15. IMPLICATIONS

The amount of variation among the interviewers has implications for interviewer training. There is, first and foremost, a need for awareness-building among the teachers who act as interviewers during the national examination interview with regard to the difference between teaching and testing practices. Many of the aspects of conducting a high-stakes oral proficiency test are new to the interviewers in Estonia. Although conducting national examinations has a history of over a decade, there are many teachers who are new to the interviewing practice. Thus, maintaining standards, rigorous time-keeping and using a script during the interview is to a very large extent of interviewers a novel idea. This is why training should play a central role in the national examination preparation procedure. Regular in-service training courses should be obligatory for all those who wish to act as interviewers in order to gain and maintain interviewer proficiency. The training should focus on the following aspects of conducting an oral interview (all of which were found lacking in the current study):

- The reasons for and practice of time-keeping during the oral interview.
- The reasons for providing a script for an interviewer.
- Practical script application.
- Monologue management vs. role-play management.

- Interviewer language (directness, hedging, politeness, power adjustment).
- Interviewer language proficiency (especially pronunciation and grammar).
- Managing the physical conditions of the oral interview (recording, background, etc.).

The study also has implications for language testing practices. Although training is of fundamental importance, that is only part of the process of maintaining interviewer standards and obtaining valid testing results. Another essential element in the process is that of monitoring interviewer practices. It is crucial for the examination procedure to be monitored by the respective specialists either from the national Examination and Qualification Centre or the Ministry of Education to make sure that the national examination is administered similarly in all schools and that the candidates who take the examination are all subjected to similar conditions irrespective of the school they attend or the interviewer they may have. The first step towards providing similar conditions to all candidates is to record all the interviews so that regular monitoring of interviewers could be conducted. Recording the interview will motivate the interviewer to adhere to the requirements more rigorously, which in turn will mean fairer testing conditions to all candidates.

The study also has research implications. The current study has attempted to isolate and pinpoint particular interviewer behavioural patterns and practices during the oral proficiency interview and has done so without recourse to how particular behaviours actually affect the candidate rating. Study of if and to what extent interviewer behaviour affects candidate's rating at the end of the interview would be a further step in the oral proficiency interview validation process. This, in the Estonian context would mean studying the interplay of the candidate, the task, the interviewer and the rater.

Further research should also be conducted observing the same interviewer during a number of oral interviews. This should include both male and female interviewers and interviewers of both Estonian and Russian background. This would allow us to make more substantiated generalisations about interviewer behavioural patterns, design more informed training courses and provide more grounded feedback to interviewers.

CONCLUSION

RESEARCH HYPOTHESES REVISITED

The current dissertation has had two distinct foci. On the more general level, it has looked at the process of developing the English language national examination paper into the nationally recognised proficiency evaluation tool that it is today. The first research question was formulated as a research hypotheses as follows:

Hypothesis 1. The current national examination in the English language will allow valid and reliable evaluation of students' language proficiency with the speaking test containing the greatest validity threat.

The data and the discussion found in chapter two seem to warrant an overall agreement with the claim of the hypothesis above. The national examination in the English language allows valid and reliable evaluation of the candidates' language proficiency in that it resorts to a language testing framework that represents a recognised construct of what constitutes foreign language proficiency. It is a skills-based testing system that utilises multiple tasks to measure candidates' reading, writing, speaking, listening and use of language structures ability. The examination papers are developed according to a uniform procedure relying on the national curriculum and test specifications, and the task quality is monitored through a pre-testing system. Examination papers contain tasks that are both objectively and subjectively marked, whereas subjective marking is conducted through the use of rating scales and multiple marking to ensure consistency of marking procedures. Qualification procedures are in place for those marking both the writing and the speaking section of the examination as well as for the interviewers of the OPI. Procedural uniformity within the speaking test is attempted through the use of interviewer scripts. Statistical analysis is used to monitor the quality of the task both during pre-testing and more thoroughly after the examination's complete administration. The examination results display a reasonable amount of consistency within and between reading, writing, listening and use of language structures section, which also testifies to the examination's reliability as a proficiency assessment. These are some of the most important features to support the current examination's validity claim.

There are, however, some features that seem to undermine the English language national examination validity. One feature pertains to the nature of tasks chosen to evaluate particular skills. Although the tasks utilised for measuring writing, reading, listening and use of language structures are valid per se, there seems to be a tendency to choose the tasks that are convenient, 'always used', rather than appropriate, though perhaps more challenging to design. Overuse of particular task types increases the role of method-effect, advantaging some and disadvantaging other test takers. A wider,

more appropriate choice of task types would alter the examination washback effect and enhance the overall examination validity.

The other, more serious challenge to the English language national examination validity is the absence of a procedure to monitor and second mark the speaking section of the national examination. Test validation is all about providing evidence for different instances of validity emerging during the examination. Though both interviewers and assessors of the speaking part have prescribed procedures to follow, their execution is left to the individual examiner and assessor. There is little room for monitoring the interviewing and assessing procedure as the OPI is only recorded if the candidate requests it. There is reason to believe, though, that a more systematic procedure for monitoring the speaking test should be in place. It is warranted by the difference often emerging in the examination results of the speaking section and all other sections, and corroborated by the findings of the current dissertation concerning the interviewer behaviour during the interview which in the majority of cases fails to provide uniform conditions to all candidates.

This takes us to the second focus of this research – interviewer behaviour during the speaking section of the English language national examination in Estonia. This problem was investigated in light of three research hypotheses.

Hypothesis 2. Interviewers conducting the OPI during the national examination in the English language will vary in their understanding of the expectations to their own and student behaviour during the oral proficiency interview.

Expectations to interviewer behaviour during the English language national examination were formalised in a new way as of the academic year 2008 in that starting with that year's national examination, all interviewers were expected to conduct the interview in a highly standardised way, following interviewer scripts. After familiarising the interviewers with the scripts and training them to use them, the current research set out to investigate the interviewer perception of their own behaviour in light of the new procedure. A questionnaire study sought the interviewers' opinion concerning the following points: their preparedness level to conduct OPIs, amount and quality of training received, usefulness of an interview script, time-keeping effort, student behaviour during the interview, the quality of speaking tasks, the marking scale, their own anxiety level and practices concerning recording the interview and examination room set-up. Cluster analysis of the results obtained revealed two broad groups of interviewers who differed from each other in their attitudes towards the interviewing process as well as in their reported interviewing practices. Group one reported problems with many aspects of the interviewing procedure (preparedness level, the amount of training received, time management, language of the script, etc.) and the members of the group modified their behaviour and language from one candidate to the next depending on the situation as they perceived it. Their behaviour seemed to be that of an accommodating language teacher rather than a consistent language tester. Group two reported few problems with different aspects of the interview

as a whole but within the group there appeared to be two subgroups whose behaviour displayed slightly varying patterns of behaviour. Subgroup one included interviewers who were faithful to the examination procedure and allegedly made no alterations to it. But they did display a level of nervousness about their interviews being recorded, which may be an indication of low confidence in their abilities as language testers and reluctance to be monitored for fear of not living up to the expected standards. Subgroup two consisted of those respondents who seemed to be the most confident about the testing procedure and their own testing practice: they reported actively seeking information concerning the national examination at the official examination web-site, finding it easy to follow the prescribed procedure, adhere to the given script and record their student performances. These findings seem to indicate that despite the attempt to achieve the opposite, at least part of the results of the OPI have been obtained under varying circumstances, with the interviewers admitting to chancing the interviewing conditions consciously depending on the needs of the situation as they see them. This would be considered a threat to the test validity.

Hypothesis 3. Interviewer language and behaviour during the oral proficiency interviews will display a high degree of adherence to the interviewer scripts provided for the speaking test.

In order to determine the degree of adherence of the interviewer language and behaviour to the prescribed scripts, fifty recorded national examination OPIs were transcribed and analysed for the presence or absence of the prescribed elements, addition of elements and changes made to the given scripts. Also, adherence to the required time-frame was monitored regarding the time allowed for the preparation of both speaking tasks and the time spent on completing task one.

Statistical analysis of the result obtained seems to refute the hypothesis above. There were just 20 per cent of the interviewers who adhered to the requirements of the speaking test. The overwhelming majority of the interviewers deviated from the prescribed procedure, seriously damaging the validity claim of this section of the examination. It has to be admitted, though, that there was an overall tendency to try and preserve the script demands in very broad terms (adherence to the overall structure of the interview, the number and nature of the tasks, the given topics for both task one and task two, providing guidance to the candidate through the interview, keeping the interviewer language to the minimum, etc.). This, however, is not sufficient to make a claim for the uniformity of testing conditions for all candidates. It has to be noted that there is a wide gap between what the interviewers perceive to be doing or what they theoretically know is expected of them and what they actually do during the interview. Seventy-four per cent of the questionnaire respondents reported that it was either very easy or mostly easy to follow the interview script; 45 per cent claim that they never changed the script and a further 35 per cent affirmed that they mostly did not make any changes. The actual interviewer behaviour paints a grimmer picture though with an 80 per cent deviation level.

Hypothesis 4. Deviations from the provided interviewer scripts will display patterns.

Comparison of the actual interviewer language as it appeared in the OPI transcripts with that of the interviewer scripts revealed additions, omissions and changes. These were detected in all sections of the OPI – introduction, task one and task two management – and did indeed display patterns. The deviations seemed to have been motivated either by a lack of familiarity with the script, the attempt to vary the otherwise tedious repetition of the scrip language from one candidate to the next, or the attempt to provide further assistance to the candidate. The interviewers employed accommodating techniques for clarification purposes such as reiteration, over-enunciation and slowing down of the tempo of their speech. They increased the accessibility of prompts by resorting to paraphrase, additional questions, allowing prolonged student commentary on the topic of their choice, and finishing student turns. Interviewers used particular techniques to establish rapport, signal dominance and vary the level of directness in their interaction. Certain features of the interviewer behaviour seemed to be culture-specific (use of candidate's first name, level of directness) or gender-specific (time management, backchannelling) or depend on the school-type (language accuracy, interviewer confidence). There were also a number of behaviours that unfortunately testified to a somewhat substandard level of interviewer proficiency (giving misleading information, correcting mistakes, script dependence, inability to perform the different roles during the interview, etc.)

IMPLICATIONS OF THE CURRENT RESEARCH

The current doctoral dissertation is the first attempt in Estonia to systematically investigate the English language national examination development in general and the functioning of its speaking part in particular. The results of the study should be considered by policy-makers in Estonia in their assessment with regard to how efficient it is as a language proficiency measurement tool at the moment and what steps ought to be taken in order to increase its efficiency.

The findings of the current research suggest a number of immediate implications. One, that a proper procedure be set up to monitor the conduct of oral interviews. As a minimum, this would mean making the recording of the interview mandatory and creating a system for second marking the OPI. Knowing that the interview could and will be listened to by somebody else (a second marker, a monitor of the interviewing procedure from the NEQC) will hopefully discipline the interviewer to follow the interviewer script more faithfully. Second-marking the interview will increase rater-reliability. Both procedures would work towards increasing speaking test validity.

A major implication of the current investigation is that being an English teaching specialist does not automatically make one a connoisseur of language testing, much less a successful language testing practitioner.

Working in a context, where language proficiency testing has become part and parcel of many students' foreign language learning experience, the teacher needs to have an understanding of the principles of modern language testing. Being a language testing practitioner for a high stakes language testing system presupposes being able to conform to much higher standards as a language tester. In order to establish and maintain those standards, it is important to considerably extend/improve the interviewer and rater training system to assist those teachers who want to qualify as interviewers.

Although task development was not an aim of the current research, the analysis of the national examination interviews indicated a need for a closer observation of the quality of the tasks included in the speaking test. More specifically, it was the second OPI task – the role-play – that seemed to serve its purpose only partially. Rather than being a role-play – allowing a more versatile display of the candidate's speaking ability – the second task seems to have been reduced to a student-initiated question and answer session. Efforts are needed here that the task would indeed allow the candidate to display language competence on a B2 level, as required by the national curriculum and the exam specifications.

In addition to the implications of each particular study discussed in the relevant chapters, current research seems to indicate a need for a more widespread study into aspects of proficiency testing in Estonia in general and the English language national examination in particular. Being a high-stakes exam, its results are considered while making a variety of important decisions: gate-keeping at various educational establishments, employment of candidates, judgements about quality of education at particular schools, quality of teaching of particular teachers, etc. For the decisions to be right, every attempt should be made to make the national examination results valid. For that purpose, research into aspects of the national examination development, administration, analysis of its results and its impact is of crucial importance.

There is a need to monitor the process of the national examination development not just within the cycle of one particular variant of the examination paper, but all across the span of the examination's life cycle. This will help the developers to ensure its quality as a measurement tool in terms of versatility in tasks and topics and the level of representation of the test specifications. At the same time, such research efforts verify that the examination continues to perform at the level that it is expected to perform, that there would be no significant alterations to the examination's complexity level.

Testing theories and practices worldwide are constantly in a state of change, which should also prompt research in Estonia to investigate the advances made in different aspects of measuring language proficiency and implement the findings in the practice of language proficiency testing within the framework of national examinations. As Fulcher and Davidson (2009) point out, it is important to 'recognise the need for both continuity and innovation' (2009:141).

DIRECTIONS FOR FURTHER RESEARCH

As there is very little research available about the different aspects of the English language national examination, suggestions made in connection with that may concern virtually any aspect of it. Some of the more important ones could be:

- Compare the Estonian English language national examination results with other validated English language proficiency exams to determine its concurrent validity.
- Investigate the rater behaviour during the writing and speaking test evaluation to determine what their decisions are guided by, to discover behavioural patterns.
- Investigate test-taker characteristics and the strategies they use during the oral proficiency interview.
- Investigate different tasks in respective skills sections to determine levels of task difficulty and validity of evaluation results.
- Research assets and drawbacks of computer-based tests versus paper-based tests to predict problems that might emerge if national examinations were to become computer-based, plans for which are already being made.
- Investigate how pass-marks are established and papers are assigned to particular criterial levels while marking the writing paper of the national examination.
- Washback and impact of the English language national examination on English language instruction.

Research into the problems above would help the English language national examination developers make more informed decisions about the respective areas of examination design while developing valid proficiency measurement instruments.

SUULISE KEELEPÄDEVUSTESTI EKSAMINEERIJA INGLISE KEELE RIIGIEKSAMI VALIIDSUSE MÄÄRAJANA. KOKKUVÕTE

SISSEJUHATUS

Uurimistöö eesmärgid ja hüpoteesid

Käesoleva doktoritöö inspiratsiooniallikaks on Eesti gümnaasiumides kehtestatud riigieksamite süsteem, mille loomine algas 1990. aastate alguses ning mis käivitus ametlikult 1997.aastal. Doktoritöö vaatleb eksamitöösse tehtud muudatuste põhjal ühelt poolt inglise keele riigieksami koostamise printsiipide kujunemist ja nende muutumist kümnendi jooksul (1997–2008). Teiselt poolt uurib doktoritöö riigieksami tulemuste valiidsust eksami suulise osa eksamineerija tegevuse funktsioonina.

Uurimistöö eesmärgid võib sõnastada järgmiste hüpoteesidena:

Hüpotees 1. Praegune inglise keele riigieksam võimaldab õpilaste keelepädevuse valiidset ja usaldusväärset hindamist, kusjuures suurimal määral ohustab testitulemuste valiidsust suulise osaoskuse test.

Hüpotees 2. Eksamineerijad, kes viivad läbi inglise keele riigieksami suulise kõneoskuse mõõtmise intervjuu, mõistavad erinevalt suulise kõneoskuse intervjuu käigus eksamineerijale ja eksamisooritajale seatud ootusi.

Hüpotees 3. Suulise kõneoskuse intervjuu läbiviimisel langevad eksamineerija keelekasutus ning tema üldine käitumine suurel määral kokku antud eksami läbiviimise käsikirjas pakutuga.

Hüpotees 4. Eksamineerija käsikirjast kõrvalekaldumistes avalduvad seaduspärasused.

Uurimismeetodid

Esimese hüpoteesi paikapidavust kontrollitakse deskriptiivseid, analüütilisi ja kontrastiivseid uurimismeetodeid kasutades, teise hüpoteesi kontrollimisel rakendatakse kõigepealt analüütilisi uurimismeetodeid praegu kehtiva eksamisüsteemi käsitlemisel ning seejärel küsitlusuuringut, mille tulemuste üldistamiseks kasutatakse Spearmani astakkorrelatsiooni ja klasteranalüüsi. Hüpoteeside 3 ja 4 õigsust hinnatakse kvalitatiivsete andmetöötlusmeetoditega, mis on adapteeritud A. Lazartoni ja A. Browni (cf. Lazarton 2002, Brown 2005) uurimustes rakendatud konversatsioonianalüüsis.

Uurimismaterjal

Uurimistöö tulemused rajanevad allpool tabelis esitatud materjalil.

Hüpotees	Materjal
1	• Riigieksami koostamist ja läbiviimist puudutavad dokumendid (1995–2008). • Riigieksamitööd (1995–2008). • Riiklik õppekava. • Inglise keele riigieksami eristuskiri (käsiraamatud) • Ajakirjanduses ilmunud vastukaja riigieksamile. • Eksamistatistika (1995–2008).
2	• Inglise keele riigieksami suulise kõneoskuse kontrollimiseks koostatud eksamineerijate käsikirjad. • Tegevusjuhised eksamineerijatele. • 81 küsitlusuuringu vastustelehte (40 väidet ja 4 vaba vastusega küsimust). • Spearmani astakkorrelatsiooni tulemused. • Klasteranalüüsi diagram.
3 ja 4	• 50 riigieksami suulise osa intervjuud - 183 – leheküljeline intervjuude transkriptsioon. - 10 tundi 32 minutit ja 27 sekundit intervjuude lindistusi.

Dissertatsioon koosneb sissejuhatusest, neljast peatükist, kokkuvõttest, kasutatud kirjanduse loetelust ja vajalikest lisadest.

1. TEOREETILINE TAUST

1. 1. KEELEOSKUSE MÕÕTMINE

1. 1. 1. Keeleoskuse mudelid

Inglise keele kui võõrkeele testimise teooriat ja praktikat on enim mõjutanud Canale ja Swaine'i (1980), Bachmani (1990), Bachmani ja Palmer'i (1996) ning Celce-Murcia, Dörnyei ja Thurrell'i (1995), ning Euroopa keeleõppe raamdokumendi (EKR) (2001) pakutud keeleoskuse mudelid.

Canale ja Swaine'i (1980) kommunikatiivse keeleoskuse mudel koosneb kahest elemendist: kommunikatiivsest keeleoskusest kui sellisest, mis hõlmab grammatilist (grammatika, leksika, morfoloogia, süntaks, semantika, fonoloogia ja ortograafia), sotsiolingvistilist (keelekasutuse ja diskursuse reeglid) ja strateegilist keeleoskust ühelt poolt ning suhtlemist ennast teiselt poolt. Autorid eristavad kommunikatiivset keeleoskust ja ja tegelikku keelekasutust, kuigi mudeli viimast elementi pole teoorias edasi arendatud. Mudeli selline formuleerimine 1980.aastal tähendas olulisi muudatusi ka keeletestimise teoorias ja praktikas: alates sellest "peavad keeletestid sisaldama ülesandeid, mis nõuavad tegelikku keelekasutust, ega kontrolli ainult keeleteadmisi" (Fulcher & Davidson 2007:39). Fulcheri sõnul on just sellele mudelile toetudes võimalik kommunikatiivsesse keeletestimisse lülitada üksikteadmiste testimine (discreet point testing) ning samuti töötada välja kriteeriumid keeleoskuse hindamiseks keelepädevuse eri astmetel (ibid).

Bachman'i (1990:87) kommunikatiivse keeleoskuse mudel, mis ilmus kümme aastat peale ülalkirjeldatud mudelit, defineerib keeleoskuse kui oskuse, mis koosneb struktuurikompetentsist ja pragmaatilisest kompetentsist. Struktuurikompetents omakorda koosneb grammatilisest kompetentsist (sõnavara, morfoloogia, sidusus, süntaks ja fonoloogia) ja tekstitundmisest (sidusus ja retooriline struktuur). Pragmaatiline kompetents koosneb samuti kahest osast: illokutsionaarsest ja sotsiolingvistilisest kompetentsist. Kommunikatiivset keeleoskust defineeritakse viie komponendi kaudu: teadmiste struktuur, keelekompetents, strateegiline kompetents, psühhofüsioloogilised mehhanismid ja keelekasutussituatsiooni kontekst. Mudeli hilisemas versioonis asendus 'teadmiste struktuuri' kategooria 'temaatiliste teadmiste' kategooriaga, samuti defineeriti strateegilist kompetentsi kui metakognitiivsete strateegiate kogumit ning mudelisse lisati ka afektiivsed faktorid. Võrreldes Canale ja Swaine'i mudeliga on antud mudel oluliselt detailsem keeleoskuse kirjeldus pakkudes samuti mehhanismi selle kohta, kuidas keeleteadmisi suhtlemisel rakendatakse. Keeletestide koostamise seisukohast tähendab Bachman'i mudel eelkõige vajadust arvestada ülesannete koostamisel suulise ja kirjaliku kõne realiseerimisel rakendatavaid erinevaid strateegiaid.

Celce-Murcia, Dörnyei ja Thurrell'i (1995) mudel pidas eelkõige silmas õppekava arendust. Nende mudel koosneb viiest elemendist: lingvistiline,

keelekäitumise, sotsiokultuuriline, diskursuse ja strateegiline kompetents. Autorid asetavad kommunikatiivses keeleoskuses kesksele kohale diskursuskompetentsi, mis rajaneb lingvistilisel, käitumis ja sotsiokultuurilisel komponendil kuna strateegiline kompetents mõjutab mudeli kõikide komponentide vastastikust toimet.

Euroopa keeleõppe raamdokument (EKR) on praegu Euroopas ilmselt kõige enam testimist mõjutav dokument. Seda defineeritakse mitte kui mudelit vaid kui raamdokumenti, mis ei paku kommunikatiivse keeleoskuse teoreetilist käsitlust vaid pigem nende oskuste kirjeldust, mis on teooriast testimise tarvis välja valitud. Siiski võib ka raamdokumendist leida selle teoreetilise aluse lühikirjelduse (vrd. CEFR:13–16). Selle järgi koosneb kommunikatiivne keeleoskus lingvistilisest, sotsiolingvistilisest ja pragmaatilisest kompetentsist, kusjuures igaüks neist koosneb teadmistest, oskustest ja teabest (know-how). Kõigi varasemate mudelitega võrreldes on EKRs pakutu teoreetiliselt kõige vähem põhjendatud. Dokumendi väärtus seisneb tema rakendatavuses näidisena hindamisprintsiipide, hindamissüsteemide, hindamisskaalade ja juhendite väljatöötamisel ja hindamisel.

1. 1. 2. Valiidsus

Valiidsus, mida loetakse testi olulisimaks omaduseks, tähendab "määra, mis näitab, kuivõrd järeldused ja otsused, mida testi tulemuste põhjal teeme on tähendust omavad, asjakohased ja kasulikud" (APA 1985, tsiteeritud Bachman 1990:25). Keeletest peab olema koostatud nii, et see mõõdaks keeleoskust, kuid mõõtmise tulemused ei sõltuks mõõtmisveast, ega muudest faktoritest, mis mõõtmisprotsessiga kaasas käivad. Traditsioonilise valiidsusteooria sõnastasid Chronbach ja Meehl (1955), kes jagasid valiidsuse võrdlevaks valiidsuseks (mis omakorda koosneb prognoosivast ja võrdlusvaliidsusest), sisuvaliidsuseks ja konstruktivaliidsuseks. Alderson jt. (1995) jagavad valiidsuse sisemiseks (näivvaliidsus, sisuvaliidsus ja vastamisvaliidsus) ja väliseks valiidsusseks (võrdlusvaliidsus, prognoosiv valiidsus ja konstrukti valiidsus). Testi valideerimine nende teooriate raamides tähendas tõestusmaterjali kogumist kõikide nimetatud valiidsusaspektide olemasolu kohta. Muid testi omadusi – usaldusväärsus, praktilisus, mõju – vaadeldi iseseisvalt, kuid siiski lõppkokkuvõttes testi valiidsusaspektide määratlejatena.

Valiidsuse teooriat arendas oluliselt edasi Messick, kelle teooriast lähtudes defineeritakse valiidsust nüüd kui "integreeritud hindavat otsust sellest, mil määral empiiriline materjal ja teoreetilised kaalutlused toetavad nende järelduste ja tegude piisavust ja asjakohasust, mida tehakse testi tulemuste või teiste hindamismudelite põhjal" (Messick 1989:13). Messick'i valiidsuse mõõtmise maatriks koosneb kahest osast – tõestusmaterjali allikast ja tulemuste funktsioonist (testi tulemuste interpreteerimine ja testi kasutamine). Testi valideerimine selle teooria põhjal tähendaks tõestusmaterjali kogumist nende otsuste kohta, mis testi tulemuste põhjal on tehtud, või selle kohta, kuidas testi tulemusi on kasutatud (ibid).

1. 2. EKSMINEERIJA VARIATIIVSUS VALIIDSUSE DETERMINANDINA

Käesoleva uurimistöö peamine suund seondub T. McNamara (1997, 2001, 2003), A. Brown'i (2003, 2005) ja A. Lazarton'i (1996, 2002) poolt käsitletud probleemistikuga. See uurimissuund ei käsitle keeletestimist kui keeletesti käigus toimuvat keelepädevuse demonstratsiooni, vaid kui sotsiaalset tegevust, mis konstrueerib 'keelepädevuse' (McNamara 2004:339). Bachman näitab, et keelepädevust ei saa tuletada keeletesti ajal, vaid seda võib mõjutada testiarendaja konstruktikontseptsioon (Bachman1990:32) ning testimise meetod (Bachman 1990:225). McNamara rõhutab, et testisooritaja keelepädevus on mitme osaleja – testikoostaja, partneri, hindaja – koosmõju tulemus' (McNamara 2001:338), ning soovitab rakendada "diskursuse analüüsi võtteid, näitamaks, et suuline keelekasutus on oma olemuselt ühiselt konstrueeritud" (McNamara 2001:340).

1. 2. 1. Suulise keelepädevuse intervjuu kui hindamisvahend

Suulise keelpädevuse intervjuu (SKI) on laialt kasutatav kõnelemisoskuse mõõtmisvahend, milles testisooritaja osaleb koos intervjueerijaga vestluse-laadses tegevuses. Sõltumata populaarsusest, on SKI-d testiteoorias palju kritiseeritud (Bachman ja Savingion 1986, Bachman 1988, van Lier 1989, Lazarton 1992, Young ja Milanovic 1992, Young 1995), väites, et "keeleliste nähtuste hulk, mida intervjuu käigus on võimalik kontrollida, on piiratud" (Cohen 1994:262). Kontrollida on võimalik fonoloogilist, leksiko-grammatilist ning mõningaid diskursuse aspekte, kuid mitte kõiki teemasid ja teksti tüüpe, diskursuse interaktiivseid aspekte nagu näiteks keelefunktsioonid või suhtlemise struktuuri või keelekasutust muudes situatsioonides (Cohen 263).

Alates 1990. aastatest kaldub uurijate huvi eksamineerija rollile intervjueerimise protsessis. Young ja Milanovic (1992), Perret (1992), Kormos (1999), O'Sullivan (2002), Fulcher ja Reiter (2003), Luoma (2004) ja teised uurivad SKI-d just intervjueerija rollile keskendudes ning viitavad tema märgatavalt suuremale osakaalule ja mõjule testimise protsessis kui seni oli käsitletud. Kuigi eksamineerija/intervjueerija annab kõnevooru intervjuu käigus meelsasti testisooritajale ja viimane kõneleb tavaliselt poole rohkem kui intervjueerija , on viimasel siiski ainukontroll intervjuu sisu ja pikkuse üle. Intervjuu käigus on eksamineerija üldjuhul suhtlemise initsiaator ning küsitleja, testisooritajal küsimuste esitamise ja teemavaliku õigus puudub. Oma olemuselt on intervjuu seega asümmeetriline, mida tavaline vestlus ei ole. SKI muudab problemaatiliseks ka sotsiaalne distants eksamineerija ja testisooritaja vahel, ametiseisundist tulenev autoriteet, mõjuvõimu hulk, distantsist tulenev viisakusväljenduse vajadus ja kultuuriline aspekt.

1. 2. 2. Eksamineerija käitumine

Eksamineerijate käitumist käsitledes kõneldakse vastavates artiklites – Ross ja Bervick (1991), Malvern ja Richards (2002), Lazarton (1996) – kohandumusest kui "protsessist, kus suhtluses osalejate keel hakkab süstemaatiliselt sarnastuma või eristuma, s.t.

muutuma sarnasemaks või erinevaks teistest vestlusest osavõtjate keelest" (Malvern ja Richards 2002:86). Konvergentne kohandumus tuleneb soovist saavutada sotsiaalset aktsepteerimist, püüdest efektiivselt partneriga suhelda. Nii Ross ja Bervick (1991) kui Lazarton (1996) esitavad taksonoomia kohandumustüüpidest, kusjuures mõlemad taksonoomiad käsitlevad vaid konvergentset kohandumust. Malvin ja Richards rõhutavad ka divergentse kohandumuse olulisust keeletestimise kontekstis, kus vastupidiselt konvergentsele kohandumusele sunnib divergentne kohandumus "testisooritajat oma keeleoskust laiemas ulatuses demonstreerima" (Malvin ja Richards 2002:101).

1. 2. 3. Eksamineerija soo mõju

Eksamineerija soo rolli on uurinud teiste hulgas O'Loughlin (2002), McNamara (2004) ja Lumley ja Sullivan (2005). O'Loughlin'i uurimus viis järelduseni, et "meeste ja naiste vestlusstiilid on väga erinevad ja selgesti eristatavad […] naiste stiil on kollaboratiivne, kooperatiivne, sümmeetriline ja toetav, kusjuures meeste vestlusstiil on ennast kehtestav, vastastikusele koostööle mitterajanev, asümmetriline ja vestluskaaslast mittetoetav" (2002:170). Sõltumata leitud vestlusstiilide erinevustest, ei leia aga O'Loughlin olulisi erinevusi intervjueerimisstiilides, mis soolisest eripärast tuleneksid. Lumley ja O'Sullivan (2005) nendivad võimalikku seost intervjueerija soo ja teemavaliku vahel, kuid rõhutavad tulemuste ebapiisavust ja edasiste uurimuste vajadust antud teemal.

1. 2. 4. Eksamineerija professionaalne pädevus ja isikupära

Eksamineerija profesionaalse pädevuse määravad ühelt poolt tema keeleline pädevus ning teiselt poolt pädevus intervjueerijana. Vastavaid seoseid keeletestimise tulemustega on uurinud Morton jt. (1997), Mcnamara ja Lumley (1997), Brown (2003, 2005), Luoma (2004). Kui uuriti professionaalse pädevuse seost eksamitulemusega, leiti, et hindajad andsid üldjuhul kõrgemaid punkte testisooritajale siis, kui eksamineerija professionaalset pädevust ei peetud nõuetekohaseks. Luoma (2004) märgib, et kui professionaalne pädevus ei ole tavaliselt testimise protsessis probleem, siis eksamineerija suhtlemisstiil kindlasti on (2004:38). Nii Luoma, kui ka Brown (2003, 2005) on leidnud suhtlemisstiilist tulenevaid eksamitulemuste erinevusi sama eksamisooritaja puhul. Brown märgib, et "eksamineerijad erinevad üksteisest selle poolest, kuidas nad teemasid valivad, informatsiooni koguvad, mil määral nad eksamineerijat toetavad, kuidas nad ülesannet sisse juhatavad ja küsimusi esitavad" (Brown 2005:206). Morton jt. (1997) lisavad siia juurde intervjueerimisõhkkonna loomise erinevused – testisooritajate julgustamine, tagasiside andmine, viisakuspiiride kehtestamine. Intervjuu õhkkonda märkavad hindajad ja see kajastub testisooritajatele antud punktides. Brown (2005) märgib, et erinevused eksamineerija käitumises võivad mõjutada eksami konstrukti ja seega tuleb neid erinevusi arvestada eksami valideerimisprotsessis.

1. 2. 5. Eksamineerijate koolitus

Võrreldes hindajate koolitusega, on eksamineerijate koolitust ja selle mõju eksamineerijate tegevusele oluliselt vähem uuritud. Fulcher (2003) põhjendab seda teadmuse puudumisest selle kohta, millisel määral eksamitulemus võib sõltuda eksamineerija tegevusest. Enamik probleemiga tegelenud teadlastest – Alderson jt. (1995), Bachman (2003), Fulcher (2003) Luoma (2004), Brown (2005), Lazarton (1996), O'Laughlin (2002), McNamara (2004) – rõhutavad eksamineerijate koolituse vajadust ning eksamineerija tegevuse erinevust keeleõpetaja tegevusest, kuid märgivad samal ajal peaaegu olematut vastavasisulist uurimistööd ning selge arusaama puudumist sellest, milles eksmineerijate koolitus peaks seisnema. Küll aga peaks koolitus olema orienteeritud sellele, et luua "õiglased testimistingimused kõikidele testisooritajatele" (Luoma 2004:38).

2. AJALOOLINE TAUST

Süstemaatiliste üleriiklike standardtestide läbiviimine inglise keeles kommunikatiivse keeleoskuse kontrollimiseks algas koos Eestis iseseisvuse taaskehtestamisele järgnenud muutustega hariduselus. Esimesed ametlikud üleriiklikud eksamid toimusid 1997. aastal, kui riigieksamid viidi läbi nii põhikooli kui ka gümnaasiumi lõpus. Inglise keele riigieksamit valmistati ette alates 31. jaanuarist 1993, kui Hariduse ja Kultuuriministri määrus nr.6 "Põhikooli ja gümnaasiumi õpilaste järgmisesse klassi üleviimise, lõpueksamite korraldamise ja kooli lõpetamise kord" jagas eksamid riigi ja koolieksamiteks ning määras nende korraldamise korra. Sama aasta detsembrist asus tööle töögrupp, kes hakkas vastavat projekti kokku panema. 1994. aastal valminud projekti käigus korraldati baasuuring (Baseline Study), mille eesmärk oli uurida inglise keele õpetamise ja testimise olukorda Eestis. Samal ajal läbis inglise keele riigieksami töögrupp spetsiaalse testimiskoolituse Lancasteri ülikoolis. Nimetatud uuringule ja väljaõppele toetudes koostati 1994. aastal riigieksami eristuskiri ning tehti esialgne plaan koostada riigieksam kahetasemelisena (lähtudes ilmselt erinevustest koolides pakutavas inglise keele õppe mahus).

Esimene katseeksam valmis siiski ühe üldise eksamina, mida pakuti valikeksamina kõikidele koolilõpetajatele sõltumata keeleõppe mahust või intensiivsusest. Katseeksamid toimusid 1995. ja 1996. aastal, kusjuures eksamisooritajate arv kasvas 222-lt 1995.a. 1304-ni 1996. aastal. Eksam on algusest peale koosnenud viiest osast – kirjutamine, kuulamine, lugemine, keelestruktuurid, kõnelemine – kuid eri osade osatähtsus eksamitöös on aja jooksul muutunud. Katseeksamid asetasid suurema rõhu keelestruktuuride kasutamisele ja retseptiivsete oskuste kontrollimisele, samas kui produktiivsed eksami osad andsid vähem punkte. Katseeksamite eesmärk oli "eksamiülesannete raskusastme määratlemine ja need vastavusse viimine eristuskirja ja õppekava nõudmistega ning samuti eksami läbiviimise protsessi eksamitööde hindamise ja tulemuste esitamise kontrollimine ning ühtlustamine"(Kristi Mere isiklik arhiiv). Katseeksamite korraldamisega avanes Hariduse ja Kultuuriministeeriumil võimalus võrrelda osaoskuste tulemusi eksami terviktulemustega ning ka aastate kaupa. Katseeksamite kogemustest tehti järgmised järeldused:

- eksami raskusaste oli adekvaatselt määratud;
- õigustust leidis üleriikliku hindamiskomisjoni kasutamine;
- ühist hindamisskaalat rakendades saavutati usaldusväärsem ja võrreldavam tulemus, mis suurendas eksami valiidsust (eriti oluline oli see subjektiivselt hinnatud eksamiosade puhul);
- kuulamisosa raskusaste polnud piisav, samuti vajati kuulamisosa edastamiseks raadiokanalit, mis oleks kõikides koolides kuuldav;

- iga eksamiosa ülesanded peavad olema järjestatud lihtsamalt keerulisemale; koolides peab olema välisvaatleja;
- kirjalikku osa peab hindama kaks hindajat, positiivse hindepiiri määramiseks tuleb kasutada eksperte;
- suulise eksami läbiviimiseks tuleb koolitada juurde nii eksamineerijaid kui ka hindajaid;
- eksamiosade sooritamiseks ettenähtud aega tuleb täpsustada (ibid).

Esimene ametlik riigieksam 1997.aastal, mille valis 9280 abiturient, järgis katseeksamite käigus väljatöötatud struktuuri. Inglise keele riigieksamile reageeriti umbes saja artikliga riigi ajakirjanduses (NE 1997:5). Positiivseks peeti eksamitulemuste ja seega õpilaste keeleoskuse võrdlemisvõimalust, erinevate hindamismetoodikate rakendamist ning osaoskuste eraldi mõõtmist. Õpilased väärtustasid õiglast hindamist ja eri osaoskuste eraldi mõõtmisest tulenevat tulemuste suuremat usaldusväärsust, samuti seda, et riigieksam hakkas kehtima nii koolilõpueksamina kui ka ülikooli sisseastumiseksamina. Negatiivsena märgiti kommentaarides eksamitöös endas või eksami läbiviimise protsessis esinenud vigu või probleeme: eksami liigne raskus, ülesannete järjestus, trükivead (Läänemets), ebapiisav eksamiks ettevalmistav materjal (Penjam). Räägiti eksami üldiselt madalatest tulemustest (Reiman, Adamson), üldisest usalduse puudumisest välishindamise suhtes (Märja), täiendavate lisameetmete vajadusest spikerdamise vältimiseks (Kapp).

Järgnenud kümnendi jooksul (1997–2008), tehti eksamistruktuuri mõned muutused: muudeti eksamiosade järjestust, pikendati kirjutamise ja kuulamisosa sooritamiseks antud aega, täpsustati kirjalikus osas nõutud sõnade arvu. Eksamiosade kaal eksamitöös võrdsustati. Nii kirjutamise kui ka kõneoskuse hindamise skaalat on muudetud vastavalt sellele, kuidas on muutunud osaoskuse olemuse mõistmine.

Eksamitulemustest annavad ülevaate alltoodud tabelid.

Tabel 2. 1. Eksamisooritajad ja nende keskmine tulemus.

Aasta	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Õpilaste arv	9280	8769	9258	9461	8488	9311	9431	9099	9415	9590	9696
Keskmine	64,6	58,8	61,8	64,1	64,9	66,6	63,99	66,6	71,9	64,4	68,8
Std*	17,7	19,9	19,9	19,7	18,8	17,8	16,9	16,7	16,0	16,1	16,0
Maksimaalne tulemus	99	99	100	99	99	100	100	100	100	99	99
Minimaalne tulemus	8	0	0	0	5	0	0	1	1	11	5

* std = standardhälve

Tabel 2. 2. Ülevaade keskmistest tulemustest (1998–2007).

Aasta	Kirjutamine	Kuulamine	Lugemine	Keelestruktuur	Kõnelemine
1998	12,2	10,1	10,7	10,4	15,6
1999	12,4	11,2	10,9	11,8	15,7
2000	12,3	11,6	13,3	9,9	15,6
2001	11,3	14,7	12,2	11,1	14,7
2002	11,6	13,2	14,7	11,9	15,5
2003	11,5	11,9	13,5	11,0	15,8
2004	13,4	12,0	13,7	11,5	16,1
2005	13,3	12,7	15,3	13,1	16,4
2006	12,9	11,3	11,9	12,1	16,6
2007	13,1	13,1	12,5	13,1	16,9

Tulemusi võrreldes näeme, et kui kirjutamise, kuulamise, lugemise ja keelestruktuuride tulemused on lähedasel tasemel, siis kõnelemisoskuse tase on mingil põhjusel teistest osaoskustest märgatavalt kõrgem.

Tabel 2. 3. Poiste ja tüdrukute keskmised tulemused.

Aasta	1999	2000	2001	2002	2003	2004	2005	2006	2007
Poisid	60,3	61,0	63,3	66,2	63,3	65,5	71,4	65,2	69,5
Tüdrukud	62,8	63,6	64,6	66,9	64,4	67,3	72,3	63,8	68,3

Tüdrukute eksamitulemused on enamasti poiste tulemustest pisut kõrgemad, mis ühelt poolt võib näidata tüdrukute paremat keeleoskust, kuid teiselt poolt ka seda, et eksamiülesanded võivad olla koostatud nii, et tüdrukutel on neid kergem sooritada.

Tabel 2. 4. Eesti ja vene õpilaste keskmised tulemused.

Aasta	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Eesti	66,8	61,2	64,4	64,6	65,8	68,3	65,6	67,5	74,1	66,3	70,7
Vene	59,2	51,5	53,5	55,6	59,1	61,8	59,3	64,2	65,2	57,8	68,8

Eesti koolide õpilaste inglise keele riigieksami tulemused on üldiselt kõrgemad kui vene koolide õpilaste tulemused. Vahe tuleneb ilmselt sellest, et eesti koolides hakatakse inglise keelt üldjuhul õppima A-keelena ja vene õppekeelega koolides B-keelena (A-keeleks on eesti keel), seega märgatavalt hiljem.

Kokkuvõte

Inglise keele riigieksami kehtestamisele eelnes põhjalik ettevalmistusperiood. Praeguseks on eksam suhteliselt hästi toimiv, abiturientide poolt kõige sagedamini valitud valikeksam, suurim Eestis väljatöötatud üleriigiline võõrkeele pädevust testiv eksam. Eksam põhineb riiklikul õppekaval ja eksami eristuskirjal, mis eeldab, et gümnaasiumi lõpuks saavutatud keeletase inglise keeles on B2 nii nagu see on sõnastatud Euroopa keeleõppe raamdokumendis.

Riigieksami ülesandetüübid küll varieeruvad nagu ka pädevuseksamilt oodatakse, kuid ülesandekoostajad näivad eelistavat teatud liiki liiga sageli korduvaid ülesandetüüpe. See võib mõjutada eksamitöö tulemuste üldist usaldusväärsust ja valiidsust.

Võrreldes teiste osaoskustega on kõnelemisoskuse kontrollimine valiidsuse seisukohalt kõige vähem usaldatav, kuna eksamisooritust kontrollib vaid üks hindaja. Tavaliselt pole hindamistulemust ka juhusliku valiku põhjal teist korda võimalik hinnata.

3. KÕNEOSKUSE KONTROLLIMISE METOODIKA INGLISE KEELE RIIGIEKSAMIL. KÜSITLUSUURING

Käesolev peatükk käsitleb kõnelemisoskuse kontrolli valideerimisprotsessi inglise keele riigieksamil, alustades vastava osaoskuse kontrolli metoodika analüüsist ning käsitledes seejärel inglise keele riigieksami suulise osa intervjueerijate ja hindajate hulgas tehtud küsitlusuuringu tulemusi.

Uue metoodika kasutuselevõtt inglise keele riigieksami suulises osas tulenes vajadusest testimisprotsess ühtlustada, tagada eksamisooritajatele ühesugused tingimused ning vältida võimaluse korral muude osaoskuste kontrolli suulise keeleoskuse kontrolli käigus. Viimase eesmärgi jaoks vajas ühtlustamist eksamisoorituseaeg, lugemis- ja kirjutamisoskuse kaasamise määr (õieti selle viimine minimaalsele tasemele) suulise keeleoskuse kontrollimise käigus, eksamisooritajatele antava informatsiooni hulk, samuti eksamineerijate tegevus suulise eksami korraldamisel. Eksamiprotsessi standardiseerimiseks koostati järgmised dokumendid:

- eksamineerijate käsikirjad (scripts) suulise eksami kõigi kolme osa läbiviimiseks,
- uus hindamisskaala,
- tööjuhend eksamineerijatele,
- tööjuhend hindajatele,
- koolitusprogramm suulise eksami läbiviijate (eksamineerijate ja hindajate) ettevalmistamiseks.

Eksamineerijate käsikirjad koostati sõnasõnalised, kusjuures eksamineerija ülesandeks oli intervjuu läbi viia nii, et eksamisooritajale antaks kogu käsikirjas olev informatsioon muutmata kujul. Eksamineerijate käsikirju kolme eksami osa jaoks illustreerivad allpool esitatud näited.

Näidis 1. Eksamineerija käsikiri suulise eksami sissejuhatuse läbiviimiseks.

STAGE 1: Introduction (maximum 2 minutes)

Greet the candidate and ask him/her to sit down.
Ask the external candidates **if they are familiar with the procedure** / explain if necessary.
Ask the candidate **if he/she wants the interview to be recorded**.
If 'Yes', switch on the cassette recorder and state the candidate's code number.
If 'No', ask if the candidate is aware that he/she can only appeal against the result of the speaking paper if the answer is recorded.

Interviewer: **Hello. (**If the candidate does not know you, tell him/her your name**) I am your interviewer today, and this is (name), your assessor. How are you today?**
If candidate responds, 'I'm fine', proceed with ***'That's good then.'***
If candidate responds, 'Quite nervous', proceed with '***Just try to relax. You'll be fine'.***

Choose **ONE** of the following scenarios to continue (vary them equally during the day):
Interviewer: **Let's talk about the weather. Do you like the weather today? Why? What is your favourite type of weather? Why?**
Thank you.

OR

Interviewer: **Let's talk about your home. Do you live in a house or a flat? What do you like about your house/ flat? Why?**
Thank you.

OR

Interviewer: **Let's talk about photographs. Do you like to take photographs? Why? Why do people usually like to look at photographs?**
Thank you.

OR

Interviewer: **Let's talk about computers. Do you like working with a computer? Why? What do people usually use a computer for?**
Thank you.

Näidis 2. Eksamineerija käsikiri suulise eksami esimese ülesande läbiviimiseks.

STAGE 2: Task 1

Interviewer: ***Now, I would like you to speak on a topic for two minutes. Before you talk, you have 3 minutes to think about what you are going to say. You can make some notes if you wish.***
Do you understand?
Here is a pencil and some paper. [hand over pencil and paper]
Please, pick a topic. [point to the cards on the table]
What's the number of your topic?
Now you have 3 minutes.

The candidate has uninterrupted preparation time for 3 minutes. (The cassette recorder should NOT be switched off for that time)
When the time is up, stop the candidate by '***Alright. Remember, you have two minutes for speaking. I'll tell you when the time is up. Please start speaking now***.'

Allow the candidate **2 minutes** of uninterrupted monologue time.
Sample Topic:
Some people think that physical education should be on students' timetable every day. Why do you think they say that? Do you agree? Give reasons.

When the candidate has been speaking for 2 minutes, find a logical way (at the end of a sentence or thought) to stop the candidate in a natural and friendly manner.

OR

When the candidate has spoken for less than 2 minutes and it is not clear if he/she has finished, ask **'*Is that all you wanted to say?'*** or ***'Was there something else you wanted to say?'***
When the candidate has completed the monologue, continue with the questions in the script in the same order they appear (unless the candidate has already answered any of them in his/her monologue, in which case skip the question).

Interviewer: ***Thank you. Now, I would like to ask you some questions.***
1. What were your favourite subjects at school? Why?
2. How important is sport in your school?
3. Why do people like some subjects more than others?
4. Can schools prepare students for life? Give reasons.

Once the candidate has finished, mark the end of the task by
'*Thank you. Let's move on to the next task*'.

Näidis 3. Eksamineerija käsikiri suulise eksami teise ülesande läbiviimiseks.

STAGE 3: Task 2 (4–5 min.)
Interviewer: ***Here is a card with a task on it. Please read it to yourself. You have 1 minute to think about it. I'll tell you when the time is up.***

Note-taking is not allowed at this stage.
When the time is up, say ***'Could you start the role-play now.'***

Use the information in the script to answer candidate's questions.
Do not give more information than the candidate asks.
Keep your answers short and natural to oral communication.

Student's cue card
You are a journalist of a British newspaper, which is considering an article about the Pärnu Film Festival. Your interviewer is an organiser of the festival.
Ask the interviewer about
1. aim
2. time the festival started
3. organisers
4. prizes awarded
5. winner of 2006
6. time of this year's festival
At the end of the talk, say whether you think you have got enough information to write an article about the festival.

Interviewer's cue card
1. The aim of the International Documentary and Anthropology Film Festival is to learn about the culture of different ethnic groups.
2. The first festival took place 21 years ago.
3. The chief of the festival is Mark Soosaar who is assisted by many people from the Pärnu Museum of New Art.
4. The Grand Prize for the best film of the festival is a hand-woven West-Estonian blanket.
5. In 2006 the Grand Prize was awarded to Arunas Matelis from Lithuania for his film "Before Flying Back to Earth".
6. This summer the festival will take place on 1–8 July.

If the candidate does not finish the role play as required (does not give a decision at the end), ask ***'Is that all you wanted to say?'***
When the candidate has finished the role play, finish the interview by
'Thank you. This is the end of the interview.'

Switch off the cassette recorder.
Before the candidate leaves the room
• tell the candidate when the scores will be announced
• ask the candidate to sign the attendance form
• collect the candidate's notes

Ülaltoodud suulise keeleoskuse kontrollimise süsteem hakkas riigieksamite raames kehtima 2008. aasta kevadel. Et selgitada, millisel määral eksamineerijad uue intervjueerimissüsteemiga toime tulid, korraldas käesoleva töö autor eksamijärgselt eksamineerijate ja hindajate hulgas küsitluse. Küsimustik püüdis välja selgitada, millisel määral eksamineerijad enda arvates eksami läbiviimiseks valmis olid ja kuidas nad hindasid eksamieelset koolitust, kui kasulik oli käsikiri intervjuu käigus, kui hästi õnnestus eksamineerijate arvates kehtestatud ajapiiridest kinnipidamine, kuidas eksamisooritajad eksami ajal käitusid (ülesande mõistmine, selgituste vajadus, ülesandeks valmistumise kiirus, ülesannete täitmise adekvaatsus), nende arvamust eksamiülesannete kvaliteedist, hindamisskaalast ja selle kasutamise lihtsusest, intervjuude salvestamisest ja eksami keskkonnast.

Klasteranalüüsi tulemusena jagunesid küsitluses osalejad kahte rühma, millest üks sisaldas omakorda sarnaselt vastanutest koosnevaid alarühmi. Esimest eksamineerijate rühma iseloomustas mõningane ebakindlus enda kui eksamineerija/intervjueerija rolli suhtes ning teatav järjekindlusetus eksami läbiviimisel. Selle rühma liikmed märkisid, et olid unustanud ajalimiidist kinni pidada ning lubanud eksamisooritajal piiramatult aega kasutada. Nende arvates oli käsikirja sõnastus kunstlik ning nad muutsid seda. Nende intervjueeritajate poolt küsitletud eksamisooritajad vajasid eksami jooksul protseduurilisi selgitusi ja sõnaseletusi. Rühma liikmed märkisid, et vajavad rohkem koolitust ning avaldasid ka soovi sellest osa võtta. Selle rühma liikmetel näis olevat raskusi loobuda enda kui keeleõpetaja rollist ning asuda keeletestija rolli.

Teist rühma iseloomustas rahulolu uue eksamisüsteemiga ja suurem järjekindlus eksami läbiviimisel. Selles rühmas eristusid aga selgelt kaks alarühma. Esimese alarühma liikmed olid rahul kõikide eksamit puudutavate aspektidega: eksamieelne koolitus, materjalid, ülesanded, teemad ja hindamisskaala. Siinsed eksamineerijad väitsid, et olid järginud rollimängu rollikaarti sõnasõnaliselt ning mõte sellest, et eksamisooritajad võiksid eksamiintervjuu lindistamist soovida tekitas neis ärevust. Selle alarühma liikmed ei kritiseerinud, ega kommenteerinud lisaks ühtegi suulise eksami aspekti. Teine alarühm oli märgatavalt analüütilisem ja aktiivsem. Selle alarühma liikmed olid külastanud väidetavalt tihti eksamikeskuse veebilehekülge, et kontrollida uute materjalide olemasolu, olid valinud teadlikult sobiva eksamiruumi ja lindistanud oma õpilaste suulist keelt õppetunnis. Nende arvates oli ajalimiidi järgimine lihtne, samuti ka eksamineerija käsikirjast kinnipidamine. Selle alarühma liikmed väitsid, et eksamisooritajad olid saanud ülesannetest hästi aru, nad pidasid monoloogiteemasid arusaadavateks ning lisaküsimusi sobivateks.

Eksamineerijate väidetavat käitumist arvestades tuleks vähemalt esimese rühma eksamineerijate poolt küsitletud õpilaste eksamitulemused teise hindaja poolt uuesti hinnata. Arvestades Morton jt. (1997) ja McNamara ja Lumley (1997) uurimistulemusi, võib eksamineerija mittestandardne käitumine eksami ajal eksamitulemusi oluliselt mõjutada.

4. EKSAMINEERIJATE KÄITUMINE INGLISE KEELE RIIGIEKSAMI SUULISE OSA LÄBIVIIMISEL EESTIS

Kui hindajate tegevust suulise eksami hindamisprotsessis on kaua uuritud (vrd. Lado 1961, Bachman 1991, Alderson jt. 1995, Fulcher 2003), siis eksamineerijaga/ intervjueerijaga seotud uuringud on palju hilisem nähtus. Fulcher ja Davidson põhjendavad seda sellega, et "oleme nüüd palju teadlikumad sellest, et diskursus konstrueeritakse ühiselt ning seega on testisooritaja käitumine osaliselt sõltuv intervjueerija käitumisest' (Fulcher ja Davidson 2007:132). Tulenevalt vajadusest vähendada eksamineerija käitumise variatiivsust inglise keele eksami läbiviimisel Eestis, võeti alates 2008. aastast eksamil kasutusele eksamineerija käsikirjad ning enne eksamit suunati eksamineerijad vastavasisulisele väljaõppele.

Käesoleva uurimuse eesmärk on tuvastada, millisel määral eksamineerijad käsikirjas ettenähtud käitumis/kõnelemismalli järgivad ning milliseid muutusi nende keelekasutuses ja käitumises tegeliku töö käigus esineb. Selleks analüüsiti juhusliku valiku põhjal 2008.aasta riigieksami käigus lindistatud suulise eksami intervjuusid, kokku 50 intervjuud (25 eesti ja 25 vene õppekeelega koolidest). Intervjuud transkribeeriti ning kontrastiivanalüüs teostati järgmistes kategooriates: osalejate omadused, lindistuse kvaliteet, intervjuuks kulutatud üleüldine aeg, eksamineerija keelekasutus intervjuu eri etappidel (sissejuhatus, ülesannete 1 ja 2 tutvustus, üleminek ühelt eksamiosalt teisele, rollimäng, intervjuu lõpetamine) intervjuu etappide läbiviimiseks ettenähtud aja seire, erinevate ülesannete juhtimine, muud tähelepanekud. Saadud andmeid analüüsiti kvalitatiivselt, vajaduse korral samuti statistiliselt, kasutades andmetöötlussüsteemi SPSS for Windows 16 ja Microsoft Excel 2007. Selleks, et hinnata, kas tulemustes ilmneb statistiliselt olulisi seoseid koolitüübi ja küsitleja sooga kasutati hii-ruut statistikat ning intervjuu osade kestuse seost intervjueerija ja kooli tüübiga analüüsiti kasutades t-testi.

Uurimuses osalejaid iseloomustab järgmine tabel.

Tabel 4. 1. Eksmineerijate jaotus soo ja koolitüübi põhjal.

			Eksamineerija sugu		
			Naised	Mehed	Kokku
Koolitüüp	Vene õppe-keelega	Count	18	7	25
		% koolitüübis	72,0%	28,0%	100,0%
		% sooliselt	45,0%	70,0%	50,0%
	Eesti õppe-keelega	Arvuliselt	22	3	25
		% koolitüübis	88,0%	12,0%	100,0%
		% sooliselt	55,0%	30,0%	50,0%
Kokku		Arvuliselt	40	10	50
		% koolitüübis	80,0%	20,0%	100,0%
		% sooliselt	100,0%	100,0%	100,0%

Intervjuu läbiviimiseks kulutatud aeg selgub alltoodud tabelitest 4.2, 4.3a ja 4.3b.

Tabel 4. 2. Intervjuule tervikuna kulutatud aeg.

N	Kokku	46
	Puudu	4
Keskmine		12 min 7s
Standardhälve		1 min 30 s
Ulatus		6 min 4 s
Miinimum		8 min 32 s
Maksimum		14 min 36 s

Tabel 4. 3a. Intervjuu kestus eri õppekeelega koolides.

Koolitüüp	Pikim intervjuu	Lühim intervjuu
Eesti õppekeelega koolid	14 min 32 sek	9 min 48 sek
Vene õppekeelega koolid	14 min 36 sek	8 min 32 sek

Tabel 4. 3b. Intervjuu kestuse variatiivsus koolitüübiti.

	N	Keskmine	Miinimum	Maksimum	Standardhälve
Eesti õppekeelega koolid	21	12min 0 s	9min 48s	14min 32s	1min 27s
Vene õppekeelega koolid	25	12min 13s	8min 32s	14min 36s	1min 33 s

Nagu tabelitest selgus, kulus intervjuudeks keskmiselt aega 12 minutit ja 7 sekundit, märgatavalt vähem kui süsteemi katsetamise käigus. Eesti ja vene õppekeelega koole võrreldes selgub, et eesti õppekeelega koolides olid intervjuud lühemad, kuid seda ainult marginaalselt.

Eksamineerijate protseduurilist käitumist analüüsides selgus, et kõikidest eksamineerijatest järgis käsikirja muutmata kujul 20%, mis tähendab, et 80% tegid sellesse eri liiki muudatusi. Kõik peale ühe eksamineerija järgisid põhimõteliselt suulise eksami intervjuu kolme-etapilist struktuuri ega muutnud kordagi eksamiteemat. Käsikirja kasutamisel ilmnesid järgmised iseärasused:

- Eksamineerijad püüdsid üldreeglina püsida ajapiirides, kuid kõrvalekalded olid sagedased, kusjuures naised püsisid ajaraamides paremini kui mehed.
- Intervjuu tervikuna ja selle erinevad osad viidi läbi kiiremini kui protseduuri katsefaasis. See tulenes peamiselt intervjueerija lühemast kõnevoorust käsikirjas ettenähtuga võrreldes. Eksamineerijad järgisid eksamisooritaja teema- ja ajakasutust järjekindlamalt kui eksamineerijatele endale esitatud nõudeid.

- Eksamineerijad kasutasid intervjuu juhtimisel küll nõutud käsikirja elemente, kuid nende järjekord ja sõnastus varieerusid süstemaatiliselt, mille tulemusena eksamit sooritavad õpilased said eksamineerijalt erineval hulgal ja erinevasisulist informatsiooni.
- Käsikirjas tehtud muudatusi oli nelja tüüpi: elementide lisamine, elementide väljajätmine, elementide järjekorra muutmine ja elementide sõnastuse muutmine.
- Eksamineerijad kasutasid konvergentset kohandumust kohati selleks, et saavutada eksamisooritajale antavate juhiste selgus. Konvergentne kohandumus seisnes siin käsikirja mõne elemendi kordamises, küsimuste ja juhiste ülipüüdliku täpsusega häälda mises, kohatises vestlustempo märgatavas aeglustamises.
- Eksamineerijad püüdsid ülesandeid arusaadavamaks muuta lihtsustades küsimusi parafraasi abil, lõpetades eksamineerija kõnevoore ja esitades lisaküsimusi teemadel, millest eksamisooritaja näis kõnelda tahtvat.
- Eksamineerijad kasutasid üldise intervjuuõhkkonna loomiseks eri võtteid (tervitamine, käekäigu järel pärmine, tagasiside andmine) erineval määral, kusjuures mõned võtted näisid olevat kultuuriliselt (eksamisooritaja eesnime sagedane kasutamine) või sooliselt (tagasiside andmine) markeeritud.
- Eksamineerijad kasutavad keelelisi vahendeid dominantsuse väljendamiseks, kusjuures esines nii dominantsuse tõstmist kui ka langetamist intervjuu käigus.
- Eksamineerijad muutsid kohati suhtlemise modaalsust.
- Üldiselt madal eksamineerijate professionaalsus ilmes järgnevas: eksitava teabe esitamine eksamisooritajale, kõrge tekstisõltuvus (intervjuu ajal interaktsioon käsikirja mitte eksamisooritajaga), tarbetute tegevuste sooritamine (vigade parandus), suutmatus rakendada erinevat keelekasutust eksami eri osades sõltuvalt osa iseloomust, kohatine vigane keelekasutus.
- Eksamineerijate tegevuse analüüs kooli õppekeele kaupa näitas kõrgemat kindlustunnet (nii keelelist kui protseduurilist) eesti õppekeelega koolide eksamineerijate hulgas võrreldes vene õppekeelega koolide intervjueerijatega.
- Sooliselt oli antud valim sedavõrd ebavõrdne, et üldistusteks ei ole alust. Esialgse tulemusena võib märkida naisintervjueerijate suuremat käsikirjatruudust eksami läbiviimisel, samuti näisid naisintervjueerijad rohkem hoolivat positiivse eksamiõhkkonna loomise vajadusest.

KOKKUVÕTE

Käesoleva dissertatsiooni eesmärk oli uurida sissejuhatuses püstitatud hüpoteeside paikapidavust.

Hüpotees 1. Praegune inglise keele riigieksam võimaldab õpilaste keelepädevuse valiidset ja usaldusväärset hindamist, kusjuures suurimal määral ohustab testitulemuste valiidsust suulise osaoskuse test.

Peatükis kaks toodud andmed ning nende analüüs annavad alust nõustuda ülalesitatud hüpoteesiga. Inglise keele riigieksam võimaldab õpilaste keelepädevust valiidselt ja usaldusväärselt mõõta, kuna see kasutab mõõtmisvahendit, mis rajaneb tunnustatud võõrkeelepädevuse konstruktile. See on osaoskuste mõõtmisel põhinev testimissüsteem, mis kasutab õpilaste lugemis-, kirjutamis-, kõnelemis- ja kuulamisoskuse ning keelestruktuuride tundmise mõõtmiseks mitmeid erinevaid ülesandeid. Eksamitöö koostatakse ühetaolise, korduva protsessi tulemusena, toetudes riiklikule õppekavale ja eksami eristuskirjale, kusjures ülesannete kvaliteeti kontrollitakse eeltestimise kaudu. Eksamitöös on nii objektiivselt kui ka subjektiivselt hinnatud ülesandeid ning subjektiivne hindamine rajaneb hindamisskaaladele ja hindamise ühtluse eesmärgil ka korduvhindamisele. Sisse on seatud kvalifitseerumisprotseduur nii kirjaliku kui ka suulise eksamiosa hindajatele ning suulise osa eksamineerijatele. Protseduurilise ühtsuse tagamiseks kasutavad suulise osa eksamineerijad intervjueerija käsikirju. Statistilist analüüsi kasutatakse ülesande kvaliteedi jälgimiseks nii eeltestimise käigus kui ka eksamitöö postvalideerimisel. Eksamitulemused on ühtlased lugemis-, kirjutamis-, kuulamis- ja keelestruktuuride osas.

Mõned inglise keele riigieksami omadused seavad küsitavusse selle tulemuste valiidsuse. Üks neist puudutab riigieksami ülesannete valikut. Kuigi eri osaoskuste testimiseks kasutatavad ülesanded on iseenesest sobivad, tundub mõnikord, et ülesande tüüpe valitakse mugavuse mitte asjakohasuse järgi. Ühe ülesandetüübi ülisagedane kasutamine viib meetodi mõju rolli kasvamisele testimise protsessis, soodustades ühtede ja takistades teiste eksamisooritajate keeleoskuse adekvaatset mõõtmist.

Teine, tõsisem küsitavus inglise keele riigieksami valiidsuse hindamisel on sellise mehanismi puudumine, mis lubaks kontrollida suulise eksamiosa läbiviimise protseduuri ja rakendada selle eksamiosa puhul süstemaatiliselt korduvhindamist. Eksami valideerimiseks tuleb esitada tõendeid valiidsuse olemasolu kohta. Kuigi nii suulise osa eksamineerijatele kui ka hindajatele on väljatöötatud detailsed eksami läbiviimise protseduurid, jääb nende järgimine konkreetse eksamineerija või hindaja hooleks. Suulise eksami tulemuste erinevus teiste eksamiosade tulemustest ja käesolevas dissertatsioonis käsitletud eksamineeritate käitumise variatiivsus suulise eksamiosa läbiviimisel näib viitavat süstemaatilisema kontrollisüsteemi vajadusele suulise eksami ajal.

Hüpotees 2. Eksamineerijad, kes viivad läbi inglise keele riigieksami suulise kõneoskuse mõõtmise intervjuu, mõistavad erinevalt suulise kõneoskuse intervjuu käigus eksamineerijale ja eksamisooritajale seatud ootusi.

Eksamineerijatele esitatud nõudmised fikseeriti uuendatud kujul 2008. aasta riigieksamiks. Siitpeale nõutakse eksamineerijalt intervjueerija käsikirja ranget järgimist. Eksamineerijate ja hindajate hulgas pärast eksamit läbiviidud küsitlus uuris eksamineerijate käsitust endale ja eksamisooritajatele esitatud nõuetest riigieksami suulise osa läbiviimise protsessis. Tulemuste klasteranalüüs osutas kahe eksamineerijate rühma olemasolule, kes erinesid üksteisest nii oma suhtumiselt eksamineerimisprotsessi kui ka praktilise käitumise poolest intervjuu käigus. Tulemused näivad viitavat sellele, et hoolimata püüdlustest saavutada vastupidist, saadi vähemalt osa suulise keelepädevuse intervjuu tulemustest olukorras, kus eksamisooritajad olid erinevas olukorras ning eksamineerijad tunnistasid, et muutsid teadlikult eksamineerimistingimusi, sõltuvalt enda hinnangust olukorrale. See vähendab eksami valiidsust.

Hüpotees 3. Suulise kõneoskuse intervjuu läbiviimisel langevad eksamineerija keelekasutus ning tema üldine käitumine suurel määral kokku antud eksami läbiviimise käsikirjas pakutuga.

Selleks, et hinnata, millisel määral eksamineerijad tegelikult intervjueerija käsikirjast kinni pidasid, transkribeeriti ja analüüsiti 50 riigieksami suulise osa intervjuud. Tulemuste analüüs kummutab seatud hüpoteesi: ainult 20 protsenti eksamineerijatest järgis intervjueerija käsikirja muutmata kujul, suurem osa kaldus käsikirjast kõrvale, vähendades nii antud eksamiosa valiidsust. Ilmnes küll põhimõtteline püüd käsikirjanõudeid järgida (säilitati üldine struktuur, ülesannete arv ja sisu jne.), kuid see ei olnud piisav väiteks, et eksam sooritati kõikide eksamisooritajate jaoks ühesugustes tingimustes. Küsitlusuuringu ja tegelike uuringute võrdlemisel ilmnes märgatav erinevus selle vahel, kuidas intervjueerijad oma tegevust tajuvad ja kuidas tegelikult intervjuusid läbi viies käituvad.

Hüpotees 4. Eksamineerija käsikirjast kõrvalekaldumistes avalduvad seaduspärasused.

Eksamineerijate keelelise ja protseduurilise käitumise võrdlemine intervjueerija käsikirja ootustega näitas, et eksamineerijad teevad eksamikäsikirjas nelja liiki muutusi: nad lisavad informatsiooni, jätavad mõned elemendid välja, muudavad elementide sõnastust ja muudavad elementide järjekorda. Muutused näisid tulenevat puudulikust käsikirjatundmisest, püüdest varieerida tüütuseni korduvat eksami protseduuri või soovist aidata eksamisooritajat. Eksamineerijad kasutasid erinevat liiki kohandumustaktikaid ning nende käitumine näis kohati tulenevat eksamineerija soolistest (aja juhtimine, tagasiside) või kultuurilistest (eesnime kasutamine, modaalsus) iseärasustest. Oli ka käitumismalle, mis andsid tunnistust eksamineerijate ebapiisavast professionaalsest ettevalmistusest.

Üldkokkuvõttes on käesoleva uurimistöö tulemustel tähendus nii tulevast praktilist pädevustestimist, inglise keele riigieksamitöö koostamist, eksamineerijate ja testikoostajate koolitust kui ka testimisalase teoreetilise uurimistöö vajadust silmas pidades.

REFERENCES

ALAS, E. 2007. Developing the National Examination in the English Language. – *Open!* 32, 2–5.

ALAS, E. 2010. Interviewer Variability in Oral Proficiency Interviews. In Nordquist, R. (ed.) *Crossing Boundaries: Studies in English Language, Literature and Culture in a Global Environment*; 9–35. Peter Lang, Internationaler Verlag der Wissenschaften.

ALAS, E., LIIV, S. 2009. Constraints of Measuring Language Proficiency in Estonia: The National Examination in the English Language. H.Metslang, M.Langemets, M.-M. Sepper (Toim.). Eesti Rakenduslingvistika Ühingu aastaraamat 5, Estonian Papers in Applied Linguistics 5; 19–32. Tallinn: Eesti Keele Sihtasutus

ALDERSON, J. C., CLAPHAM, C., WALL, D. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

AMERICAN Psychological Association. 1985. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

BACHMAN, L. F. , SAVIGNON, S. J. 1986. The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. – *The Modern Language Journal* 70, 4:380–90.

BACHMAN, L. F. 1988. Problems in examining the validitiy of the ACTFL oral proficiency interview. – *Studies in the Second Languag Acquisition* 10, 149–64.

BACHMAN, L. F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

BACHMAN, L. F., PALMER, A. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

BACHMAN, L. 2000. Modern language testing at the turn of the century: assuring that what we count counts. – *Language Testing* 17, 1–42.

BACHMAN, L. F. 2002. Some Reflections on Task-Based language performance Assessment. – *Language Testing* 19; 453–76.

BAGARIĆ, V., MICHALJECIĆ-DJIGUNOVIĆ, J. 2007. Defining Communicative Competence. – *METODIKA*: Vol.8, br.14 (1/2007), 94–103.

BERRY, V. 2007. *Personality Differences and Oral Test Performance*. Peter Lang. Europäischer Verlag der Wissenschaften.

BERWICK, R.., ROSS, S. 1996. Cross-cultural pragmatics in oral proficiency interview strategies. In. Milanovic, M., SAVILLE, N. (eds.) *Performance Testing, Cognition and Assessment. Selected papers from the 15th Language testing research Colloquium.* Cambridge: Cambridge University Press, 34–54.

BROWN, A. 2003. Interviewer Variation and the Co-Construction of Speaking Proficiency. *Language Testing*, 20, 1–25.

BROWN, A.., McNAMARA, T. 2004. 'The devil is in the detail': Researching gender issues in language assessment. – *TESOL Quarterly*, 38, 524–38.

BROWN, A. 2005. *Interviewer Variability in Oral Proficiency Interviews*. Peter Lang. Europäischer Verlag der Wissenschaften.

BROWN J. D., HUDSON T. 2002. *Criterion-Referenced Language Testing*. Cambridge Applied Linguistics. Cambridge: Cambridge University Press.

BROWN J.D., ROGERS, T. S. 2002. *Doing Second Language Research*. Oxford: Oxford University Press.

BUTLER, J. 1990. *Gender Trouble: Feminism and the Subversion of Identity.* New York and London: Routledge.

BUTLER, J. 1993. *Bodies That Matter: on the Discursive Limits of Sex*. New York and London: Routledge.

CHALHOUB – DEVILLE, M. 1997. Theoretical models, assessment frameworks and test construction. – *Language Testing* 14, 1, 3–22.

CANALE, M., SWAIN, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. – *Applied Linguistics* 1, 1–47.

CANALE, M. 1993. On some dimentions of language proficiency. In Oller, J. W. (ed.) *Issues in Language Testing Research*. Rowley, MA: Newbury House, 333–42.

CEFR 2001 = Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching and Assessment 2001*. Cambridge: Cambridge University Press. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

CELCE-MURSIA, M., D. LARSEN-FREEMAN , H. WILLIS 1999. *The Grammar Book*. An ESL/EFL Teacher's Course. Heinle&Heinle Publishers.

CELCE-MURSIA, M., DÖRNYEI, Z., THURELL, S. 1995. Communicative competence: apedagogically motivated model with content specifications. – *Issues in Applied Linguistics* 2, 5–35.

CRONBACH, L. J. and MEEHL, P. E. 1955. Construct validity in psychological tests. – *Psychological Bulletin*, 52, 281–302.

CSÉPES, I., EGYÜD G. *Into Europe: The Speaking Handbook*. http:// www.lanc.ac.uk/fass/projects/examreform/into_europe/speaking.pdf

COHEN, A.D. 1994. Assessing Language Ability in the Classroom. Heinle & Heinle Publishers.

CURRICULUM 2002 = Põhikooli ja Gümnaasiumi Riiklik Õppekava 2002. Riigi Teataja I, Nr.20, Tallinn.

DAVIES, A., BROWN, A., ELDER, D., HILL, K., LUMLEY, T., McNAMARA, T. 1999. Dictionary of Language Testing. *Studies in Language Testing* 7. Cambridge: Cambridge University Press.

DAVIES, L. 2009. The influence of Interlocutor Proficiency in a paired Oral Assessment. – *Language Testing* 2009; 26; 367–396.

ERELT, T., MERE, K., PÄRN, H., SIMM, L., TÜRK, Ü. 2003. *Testiterminite seletussõnastik*. Haridus- ja Teadusministeerium, keeletalitus, Tallinn – Tartu.

FULCHER, G. 2003. *Testing Second Language Speaking*. London: Longman/Pearson Education.

FULCHER, G., REITER, R.M. 2003. Task Difficulty in Speaking Tests. – *Language Testing,* 20; 321–44.

FULCHER, G., DAVIDSON, F. 2007. *Language Testing and Assessment. An Advanced Resource Book*. London and New York: Routledge Applied Linguistics.

FULCHER, G., DAVIDSON, F. 2009. Test Architecture, Test Retrofit. *Language Testing* 26, 123–144.

GLOVER, J. A., BRUNING, R.H. 1987. *Educational Phychology. Principles and Applications*. Little, Brown and Company. Boston/Toronto.

GYMES, D. 1972. On Communicative Competence. In Pride, J.B. and Holmes, J. (eds.) *Sociolinguistics: Selected Readings*. Harmondsworth: Penguin, 269–93.

HAMP-LYONS, L. 1990. Second language writing:assessment issues. In: Kroll, B. (Ed.) *Second language writing. Research insights for the classroom.* Cambridge: CambridgeUniversity Press.

HAMP-LYONS, L. (Ed.) 1991. *Assessing Second Language Writing in Academic Contexts*. Ablex Publishing Corporation. Norwood in Jersey.

HAUSENBERG, A.-R., KIKERPILL, T., RÕIGAS, M., TÜRK, Ü. 2004. *Keeleoskuse mõõtmine*. TEA Kirjastus, Tallinn.

HOMBURG, T.J. 1984. Holistic evaluation of ESL compositions: Can it be validated objectively? – *TESOL Quarterly*, 18(1), 87–107.

HUGHES, A. 1989. *Testing for Language Teachers*. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

HUGHES, A. 2003. *Testing for Language Teachers*. 2nd ed. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

IELTS http://www.ielts.org/

JUHANSON, A. 2007.Taas kord inglise keele riigieksamist. http://www.opleht.ee/Arhiiv/2007/25.05.07/aine/3.shtml

JÕUL, M., TÜRK, Ü. 2002. Häda riigieksami pärast. http://www.opleht.ee/Arhiiv/2002/06.12.02/tekstid/dialoog/2.html

JÕUL, M., LÄTT, V., MERE, K., SASS, E., TÜRK, Ü., VILU, M. 2005. *Year 12 Handbook*. 2005. Tallinn: Argos.

KORMOS, J. 1999. Simulating Conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. – *Language Testing* 16, 163–88.

KAPP, P. Riigieksamitest, palgast, maa soolast. 1997.
http://greta.cs.ioc.ee/~opleht/Arhiiv/97Mai30/artikkel4.html

KROLL, B. (Ed.) 1990. *Second language writing. Research insights for the classroom*. Cambridge: Cambridge University Press.

LADO, R. 1961. *Language Testing*. London: Longman.

LAZARTON, A. 1992. The structural organisation of a language interview: a conversation analytic perspective. – *System* 20, 373–86.

LAZARTON, A. 1996. Interlocutor support in oral proficiency interviews: the case of CASE. – *Language Testing* 13, 151–72.

LAZARTON, A. 2002. A Qualitative Approach to the Validation of Oral Language Tests. Studies in *Language Testing* 14. Cambridge: Cambridge University Press.

LIIMAL, P. 1997.Riigieksamid läksid rahuldavalt.http://www.epl.ee/artikkel/20843

LIIV, S. 2002. Foreign Language Competence and Testing. − Liiv, Suliko (ed.), *Perspectives on English and American Language and Literature*. Tallinn: TPÜ Kirjastus, 51–59.

LIIV, S., ALAS, E. 2009. Evaluation de l'anglais dans un examen national: résultats et problème. in *Synergies Pays riverains de la Baltique*, numéro 6 "Problématiques culturelles dans l'enseignement-apprentissage des langues-cultures, mondialisation et individualisation : approche interdisciplinaire", Gerflint, France, (241–247).

LUOMA, S. *Assessing Speaking*. Cambridge Language Assessment Series. Cambridge: Cambridge University Press 2004.

LUMLEY, T., O'SULLIVAN, B. 2005. The Effect of Test-taker Gender, Audience and Topic on Task Performance in Tape-Mediated Assessment of Speaking. – *Language Testing,* 22, 415–36.

LUMLEY, T., McNAMARA, T.F. 1995. Rater Characteristics and Rater Bias: Implications for Training. – *Language Testing* 12, 54–71.

LUNZ, M.E., WRIGHT, B.D., LINCARE, J.M. 1990. Measuring the Impact of Judge severity on Examination Scores. – *Applied measurement in Education*, 3, 331–45.

LÄTT, V., MERE, K., SASS, E., TRUUS, K., TÜRK, Ü. *Year 12 Project. Estonia. Working Papers.* Lancaster, 14 December 1994.

LÄÄNEMENTS, U. 1997. Inglise keele riigieksam 1997 – muljeid ja mõtteid. http://greta.cs.ioc.ee/~opleht/Arhiiv/97Jun20/artikkel8.html

MALMBERG, K. 19. 03. 1997. Riigieksamid said seaduslikuks. – *Eesti Päevaleht*. http://www.epl.ee/artikkel/16994

MALVERTN, D., RICHARDS, B. 2002. Investigating Accommodation in Language Proficiency Interviews Using a New measure of Lexical Diversity. – *Language Testing,* 19, 85–104.

MATERIALS for the Guidance of the Test Item Writers.
http://www.alte.org/downloads/index.php

McNAMARA, T.F. 1996. *Measuring Second Language Performance*. London: Longman.

McNAMARA, T.F. 1997. Interaction in Second language performance assessment: Whose performance? – *Applied Linguistics* 18, 4, 446–465.

McNAMARA, T. 2000. *Language Testing*. Oxford: Oxford University Press.

McNAMARA, T. 2001. Language Assessment as Social Practice: Challenges for Research. – *Language Testing,* 18, 333–349.

McNAMARA, T. F., LUMLEY, T. 1997. The Effect of Interlocutor and Assessment Mode Variables in Overseas Assessment of Speaking Skills in Occupational Settings. – *Language Testing*, 14, 2, 140–56.

MESSICK, S. 1989. Validity. In. Linn, R.L. (ed.) *Educational Measurement*. New York: Macmillan/American Council on Education, 13–103.

MORTON, J., WIGGLEWORTH, G. and WILLIAMS, D. 1997. Approaches to the evaluation of the interviewer performance in oral interaction tests. In Brindley, G. and Wiggleworth, G, (eds.) *Access:Issues in English Language Test Design and Delivery*, Sidney: National Centre for English Language Teaching and Research, 175–96.

NE 1997 = Inglise keel. Riigieksam 1997. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 1998 = Inglise keel. Riigieksam 1998. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 1999 = Inglise keel. Riigieksam 1999. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2000 = Inglise keel. Riigieksam 2000. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2001 = Inglise keel. Riigieksam 2001. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2002 = Inglise keel. Riigieksam 2002. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2003 = Inglise keel. Riigieksam 2003. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2004 = Inglise keel. Riigieksam 2004. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2005 = Inglise keel. Riigieksam 2005. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2006 = Inglise keel. Riigieksam 2006. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NE 2007 = Inglise keel. Riigieksam 2007. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus.

NEQC = Riiklik Eksami- ja Kvalifikatsioonikeskus [National Examination and Qualification Centre]. www.ekk.edu.ee (05.09.2008)

O'LOUGHLIN, K. 2002 The Impact of Gender in Oral Proficiency Testing. – *Language Testing*, 19, 169–92.

O'SULLIVAN, B. 2002. Learner Acquaintanceship and Oral Proficiency Test Pair-Task Performance. – *Language Testing*, 19, 277–95.

O'SULLIVAN, B. 2008. *Modelling Performance in Tests of Spoken Language*. Peter Lang. Europäischer Verlag der Wissenschaften.

PENJAM, T 1997. Trükivigadega harjutustest, inglise keele riigieksamist ja muustki. http://greta.cs.ioc.ee/~opleht/Arhiiv/97Mar28/artikkel13.html

PERRET, G, 1990. The language testing interview: A reappraisal. In J.H.A.L. de Jong &D.K. Stevenson (Eds.), *Individualizing the assessment of language abilities*, pp. 225–238. Clevedon, England: Multilingual Matters.

RAJANGU, V. *Ministry of Culture and Education Booklet of Educational Statistics No 4, 1994:Foreign language learning*. Ministry of Culture and Education. Tõnismägi str. 11, EE0110, Tallinn.

REED, D.J., HALLECK, G.B. 1997. Probing above the ceiling in oral interviews: what's up there? In A. Huhta, V. Korhonen, L. Kurki-Suonio and S. Luoma (eds.), *Current Developments and Alternatives in Language Assessment:Proceedings of LTRC* 96, pp. 225–38. Jyväskylä, Finland: University of Jyväskylä and University of Tampere.

REED, D.J., COHEN, A.D. 2001 Revisiting Raters and Ratings in Oral Language Assessment. – Studies in Language Testing 11:82–96. Cambridge University Press.

REGULATION 2001 = Õpitulemuste välishindamise põhimõtted, riigieksamitööde, põhikooli lõpueksamitööde ja üleriigiliste tasemetööde koostamise, hindamise ja tulemuste analüüsi alused. Haridusministri määrus (Nr. 18, 23.1.2001), Tallinn.

REIMAN, L. 1997. Pressikonverentsil räägiti riigieksamitest. http://greta.cs.ioc.ee/~opleht/Arhiiv/97Jun27/artikkel4.html

RIIGIEKSAM – kas nuhtlus või õnnistus? 1997 (õpilaste arvamused) http://greta.cs.ioc.ee/~opleht/Arhiiv/97Mar21/artikkel10.html

RIIKLIK võõrkeelte ainekava.EV Valitsuse määrus nr 228 6.11.1996.

RIIKLIK võõrkeelte ainekava EV Valitsuse määruse nr 56 25.01.2002.

ROSS, S. 1992. Accommodative Questions in Oral Proficiency Interviews. – *Language Testing* 9, 2, 173–86.

ROSS, S., BERWICK, R. 1992. The Discourse of Accommodation in Oral Proficiency Interviews. – *Studies in Second Language Acquisition*, 14, 159–76.

SWAIN, M. 2001. Examining Dialogue: another approach to content specification and to validating inferences drawn from test scores. – *Language Testing* 18, 275–302.

TANKÓ, G. Into Europe. The Writing Handbook. www.lanc.ac.uk/fass/projects/examreform/into_europe/writing.pdf

TOEFL http://www.ets.org/toefl

TÜRK, Ü. 1997. Riigieksamitest inglise keele eksami ettevalmistamise kogemuste põhjal. http://greta.cs.ioc.ee/~opleht/Arhiiv/97Mar28/artikkel12.html

UNDERHILL, N. 1987. *Testing Spoken Language*. Cambridge: Cambridge University Press.

UNT, I. 1997. Riigieksamitest mõne teise nurga alt. http://www.postimees.ee/aspseek/vanapm/97/07/14/uudis.htm#viieteistkymnes

UPSHUR, J.A., TURNER, C.A. 1999. Systematic Effects in the Rating of Second-Language Speaking Ability: test method and Learner Discourse. – *Language Testing* 16; 82, pp. 82–111.

van LIER, L. 1989. Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversations. – *TESOL Quarterly* 23, 480–508.

VAUGHAN, C. 1991. Holistic assessment: What goes on in the rater's mind? In Hamp-Lyons, L.(Ed.), 111–25.

WEIR, C., J. 1988. *Communicative Language Testing*. University of Exeter.

WEIR, C., J. 1990. *Communicative Language Testing*. New York: Prentice Hall International.

YOUNG, R.E. 1995. Conversational Styles in Language Proficiency Interviews, – *Language Learning,* 45, 1, 3–42.

YOUNG, R.E., MILANOVIC, M. 1992. Discourse variation in Oral Proficiency Interviews. – *Second Language Acquisition*, 14, 403–24.

YULE, G. 2000. *Pragmatics.* Oxford:Oxford University Press.

APPENDICES

APPENDIX 1. MARKING SCALES FOR THE NATIONAL EXAMINATION IN THE ENGLISH LANGUAGE IN ESTONIA

a) A Marking Scale for Letters

	Task Completion (TC)	**Language (L)**
4	Responds to all aspects of the prompt. Ideas presented and supported. Clear organisation (uses paragraphs, logical). Correct format.	Lexically and grammatically correct. Appropriate tone. Complex sentences. Correct spelling.
3	Responds to most aspects of the prompt. May lack support. Clear but does not have paragraphs. Mostly correct format.	Lexically and grammatically mostly correct, with few unsystematic spelling mistakes. Tone inappropriate at times.
2	Important parts of the prompt not mentioned. May require re-reading because of poor organisation. Faulty format.	Basic vocabulary and grammar well controlled. Mostly simple sentences. Inappropriate tone. Frequent grammar and spelling mistakes.
1	Attempts to write a letter but most aspects of the prompt have not been addressed. Hard to follow due to lack of organisation.	The text abounds in grammar and spelling mistakes but can still be comprehended. Writer has minimum control of the lexis and grammar.
0	Does not write a letter. The prompt has been ignored. Fewer than 50 words.	Errors in grammar and spelling predominate to the extent that the text cannot be understood.

b) A Marking Scale for Essays and Reports

	Task Completion	Organisation	Vocabulary	Grammar
3	Addresses all aspects of the prompt. Ideas are presented and supported by examples.	The message can be followed without an effort. All required elements present. Clear paragraphs. Each paragraph has one central topic, which is developed. Linking devices used within and between paragraphs.	Appropriate, wide vocabulary. Error-free word-formation. Formal register. Correct spelling.	A wide range of grammatical structures. Complex sentences predominate. Punctuation well managed.
2	Addresses the prompt partially. Does not always support.	Organisation is evident but may not always be logical. Some required elements missing. No paragraphs, but logical. Some linking devices used.	Good general control of vocabulary. Mostly correct usage. Inconsistent register. Word-formation problems. Some spelling mistakes.	Mostly error-free grammar. Simple sentences predominate. Punctuation errors.
1	The content is barely connected with the prompt. Mentions or copies the prompt without developing.	No apparent organisation. Random, illogical paragraphs. Relations between ideas unclear. Linking devices mostly not used or overused.	Vocabulary quite limited. Frequent incorrect usage. Inappropriate register. Spelling-mistakes make comprehension problematic.	Limited range of grammar with frequent errors.
0	Ignores the task Plagiarised work.	The writing does not communicate. Plagiarised work.	Misspelling prevents understanding. Plagiarised work.	No ratable language. Plagiarised work.

APPENDIX 2. GUIDELINES FOR MARKERS, EXAMINERS AND ASSESSORS

a) Guidelines for the markers of writing papers of the national examination in the English language in Estonia

I. Letters

Writing a letter constitutes the first task of the writing section of the national exam in English. Students are expected to write either a semiformal or a formal letter the length of which is 120 words. The addressee and the content of the letter are specified by the prompt/ rubric. The student can make notes, which will not count towards the final number of points given for the letter. The space allotted in the written paper for writing the short task will have a pre-written date on the form, so the student will not have to write it. The assessment of the letter is based on two broad criteria: task completion and language. The maximum number of points scored for either criterion is 4, so the maximum number of points that can be awarded for the letter is 8. The assessment should be conducted according to the National Examination Marking Scale for Letters where conditions for each score have been specified. The current document serves as commentary for the level descriptors therein.

The criterion of **task completion** (TC) comprises the content, organisation and format of the letter.

TC4. Four points can be awarded when all the points of the prompt have been adequately covered, i.e. mentioned and elaborated (commented on, briefly developed). The text has paragraphs that are logical, i.e. one idea is discussed per paragraph. The text flows smoothly. All the required parts of the letter are present (salutation, purpose for writing, body of the letter, required action, if appropriate, closing remarks, sign-off).

TC3. Three points can be awarded when most aspects (two out of three) of the prompt have been adequately (cf. TC4) covered in the letter or all three prompts have been mentioned and one of them has not been elaborated. All the points have been adequately covered but the text has no paragraphs, or the paragraphs are illogical. The format is mostly correct (one of the listed features in TC4 missing).

TC2. Two points can be awarded when two prompts out of three have not been addressed. All the prompts have been only mentioned and it is clear that they have been copied verbatim from the prompt. Although prompts have been developed the organisation of the text is not logical and requires rereading to be understood. The text has no

paragraphs, or the paragraphs are illogical. Two formatting errors (cf. list in TC4).

TC1. One point can be awarded when the letter format has been attempted but the reader has trouble connecting it to the prompt. Parts of the prompt may have been mentioned but the organisation is quite random. The text has no paragraphs, or the paragraphs are illogical. Several features (three or more) of the format are either missing or wrong.

TC0. No points are awarded for task completion if the prompt has been ignored altogether, although the student has written a letter. Also, when instead of the letter some other text-type has been produced (e.g. a description, a dialogue, a poem, a report, an essay, a fairy-tale, etc.)

The criterion of **language** (L) comprises vocabulary, grammar, spelling, punctuation and register.

L4. Four points can be awarded when the vocabulary is appropriate for the task and correctly spelt. The writer uses complex sentence structures and a variety of grammatical forms (active and passive tenses, non-finite forms). Grammatical errors are extremely rare and certainly not of a systematic nature. There are no lapses in capitalisation (English, Estonian, Tuesday, March, etc.) and basic punctuation rules (full stops at the end of affirmative sentences and indirect questions, a question mark at the end of a direct question, no comma before *that*, a comma after an if-clause, upper case inverted commas before and after a direct quote, its vs. it's). The tone is correct all through, i.e. the writer does not use colloquialisms or slang in a formal letter. The text may have 2 mistakes of any nature.

L3. Thee points can be awarded when the writer uses complex sentence structures and the text is lexically and grammatically mostly correct. There are no grammatical errors or spelling mistakes, but there are frequent capitalisation and punctuation problems. The writer uses slang or colloquialisms occasionally. The text may have up to 5 mistakes of any nature.

L2. Two points can be awarded when the writer uses simple language and does not attempt complex sentences. If they are attempted, there are frequent errors. The writer uses mostly slang words or grammar (gonna, gimme, wanna). The text may have up to 8 mistakes of any nature.

L1. One point can be awarded when the text can be understood although there are numerous grammatical errors and spelling mistakes. The text is offensive (contains swear words and vulgarisms). The text has 9 or more mistakes of any nature.

L0. No points can be awarded if the text is, for the most part, incomprehensible.

II. Essays and Reports

Writing an essay or a report constitutes the second part of the writing component of the national exam in English. In both cases, the student has to write 200 (+/ – 10) words, responding to the prompt given.

In case of an essay, the aim should be to produce a text that consists of 4–5 paragraphs discussing a specified topic. The structure of a traditional essay is fairly well established and regulated. It should have and introduction (paragraph 1) that catches the reader's attention and states the main thesis of the essay – the point of view that the writer is going to discuss. Two to three paragraphs (paragraphs 2 to 3 or 4) discuss the different aspects of the thesis statement. Each paragraph first states the topic of the paragraph in the topic sentence and then illustrates it with examples, statistics or quotes. The last paragraph makes up the conclusion, which first summarises the discussion, linking it back to the thesis statement and then adds concluding remarks.

In case of a report, the writer should similarly aim for a well-structured piece of writing responding to the prompt given. The structure of a report, too, is fairly predictable. The introduction states the aim of the report, the status of the writer and the source of the information that the report is based on. The body of the report usually consists of 2–3 paragraphs that present the findings of the survey and give examples/statistics to support the findings. The conclusion summarises the main trends of the findings and makes recommendations for further action. The report is not signed like a letter. The paragraphs may have subheadings but do not have to have them as different report formats are suggested by the text-books and consequently taught to students.

The assessment of the essay/report is conducted relying on the following criteria: task completion, organisation, vocabulary, grammar. The maximum number of points that can be scored for each criterion is 3 and the minimum is 0, so the maximum possible number of points that can be awarded for the essay/report is 12. The assessment should be conducted according to the National Examination Marking Scale for Essays/Reports where conditions for each score have been specified. The current document serves as commentary for the level descriptors therein.

The criterion of **task completion** (TC) assesses how closely the prompt content has been addressed.

TC3. Three points can be awarded when all the aspects of the prompt have been fully addressed, i.e. not copied from the prompt but appropriately adapted/paraphrased/modified and illustrated with examples or other types of supporting evidence (statistics, quotes).

TC2. Two points can be awarded when two out of three aspects of the prompt have been addressed as described in TC3, or all three aspects have been mentioned but no supporting evidence has been given with at least one of the prompts.

TC1. One point can be awarded when one out of three prompts has been fully addressed or if all the three prompts have only been mentioned without illustration.

TC0. No points can be awarded if the student does not write an essay or a report or if the writing completely ignores the prompt/rubric.

The criterion of **organisation** (O) addresses the build-up and logicality of the essay/ report.

O3. Three points can be awarded if the writing is logical, clear and understandable without any re-reading. The text has been divided into logical paragraphs (one idea per paragraph). The writer has used appropriate linking devices between paragraphs (e.g. to begin with, secondly, next, etc.) and inside them (consistent personal pronoun use, relative, interrogative, demonstrative pronoun use).

O2. Two points can be awarded when there are occasional organisation problems, e.g. the text is logical but there are no paragraphs. The text has paragraphs but they are not always logical. The text has either the introduction or the conclusion missing. The linking devices have been used mostly correctly. The writing also receives 2 points for organisation if linking devices have been overused, i.e. there is a linking device starting every sentence or one linking device (e.g. and, however, etc) has been used throughout the essay/ report.

O1. The text has noticeable organisation problems: paragraphs are missing or mechanical (the paragraph mentions several issues without apparent connection). Understanding the text requires rereading because the ideas are disconnected. Linking devices are missing or the same linking device has been overused (and, but, however, etc).

O0. No points can be awarded if the text cannot be understood because of the coherence problems.

The criterion of **vocabulary** (V) addresses the issues of word – formation, range and appropriacy.

V3. Three points can be awarded if the writer has a very good command of the vocabulary required by the curriculum on this level. The writer uses the words appropriately, keeps to the formal register, has good control of word building and makes virtually no (up to 2) spelling mistakes (including capitalisation).

V2. Two points can be awarded if the writer has good general control of the vocabulary, but may have occasional errors in spelling, collocations (e.g. *interested about, *discuss about, etc.) and word-formation (e.g. *unlogical, *teached). Also, if there are lapses into slang without any apparent need. There is only occasional inappropriate use, all in all, 5 vocabulary errors of any nature.

V1. One point can be awarded if the writer's vocabulary is limited and words are frequently misspelt and misused although the text can be comprehended. There are no more than 8 vocabulary errors of any nature.

V0. No points can be awarded when the words are misspelt to the extent that the communication is broken. Also if there are 9 or more vocabulary errors.

The criterion of **grammar** (G) addresses the issue of sentence construction and punctuation.

G3. Three points can be awarded if the writer appropriately uses predominantly complex grammatical structures (active and passive tenses, non-finite forms, etc). The basic punctuation rules are well managed (full stops at the end of affirmative sentences and indirect questions, a question mark at the end of a direct question, no comma before *that*, a comma after an if-clause, upper case inverted commas before and after a direct quote, its vs. it's). There may be 2 unsystematic grammar mistakes.

G2. Two points can be awarded if the writer has used a variety of grammar constructions but these occasionally contain errors, or in case the writer uses correct grammar but simple sentences predominate. The same tense form has been used all through. Punctuation errors. There are no more than 5 grammar mistakes of any nature.

G1. One point can be awarded if the writer mostly uses formulaic language and the forms used contain frequent errors. There are no more than 8 grammar mistakes of any nature.

G0. No points can be awarded if the writer fails to write complete sentences. There are random phrases only. The text does not communicate. Also, when there are 9 or more grammatical errors of any nature.

* Scoring zero points for one criterion does not mean automatic scoring of zero points in other criteria.

b) Guidelines for the Examiners and Assessors at the national examination in the English language in Estonia.

GUIDELINES FOR THE ORAL PART OF THE EXAMINATION 2008

The materials enclosed are the property of Examinations and Qualifications Centre and all the packages (including all the cassettes) should be returned to the local Education Authorities on the last day of the oral part of the examination in your school.
There is a separate package for each day.
The enclosed material should be used during one examination day in your school.
The material is CONFIDENTIAL and information about the content MUST NOT be made known to candidates, colleagues in your school or to other schools, or to anyone else UNTIL THE END OF THE EXAMINATION PERIOD.
At the end of each examination day, the contents of the package should be put into the envelope provided and placed in a safe.

CONTENTS OF A PACKAGE

• Guidelines for the oral part of the examination
• Interviewer's and assessor's procedures

For the interviewer

• Interviewer's script for Stage 1
• Six scripts for Stages 2 and 3
• Twelve student cards for Stage 2 – two for each script
• Six student cards for Stage 3 – one for each script
• Cassettes for recording the interviews

For the assessor

• Marking scale
• Six scripts for Stages 2 and 3
• Assessment forms (*protokollid*) for recording the candidates' scores to be signed at the end of the examination

INTERVIEW FORMAT

The interview consists of three stages:
• **Stage 1** – Introduction (not assessed) – up to 2 minutes
• **Stage 2** – Monologue and discussion – between 8 and 9 minutes
• Preparation for the monologue – 3 minutes
• Monologue – 2 minutes
• Discussion – up to 4 minutes
• **Stage 3** – Role-play
• Preparation – 1 minute
• Role-play – about 4 minutes

BEFORE STAGE 1:

The interviewer has placed the student cards for Stage 2 face down on the table.
The interviewer greets the candidate and asks him/her to sit down. The interviewer asks the candidate whether he/she wants the interview to be recorded. When the answer is 'Yes', the interviewer will switch on the cassette recorder and state the student's code number. When

the answer is 'No', the interviewer asks if the candidate is aware that he/she can only appeal against the result of the speaking paper if the answer is recorded.

STAGE 1:

The interviewer proceeds with **Script for Stage 1**. The interviewer follows the script wording without omissions or paraphrase. This part is not assessed as its aim is to relax the candidate and prepare him/her for the interview. Therefore, the phase should not last more than 2 minutes.

STAGE 2:

The interviewer follows the **Script for Stage 2**. The interviewer asks the candidate to choose one of the twelve cards lying face down on the table. The interviewer asks the candidate to say the number of his/her topic so that the assessor can write it down. The candidate then quietly reads the information on the card. The candidate has 3 minutes to prepare his/her monologue. The candidate can take notes (sheets of paper and pencils/pens should be provided) while planning his/her presentation. (The scrap paper used will stay in school but must not be taken out of the examination room and will have to be destroyed at the end of each examination day.)

When 3 minutes have passed or the candidate is ready, the interviewer asks him/her to start. The monologue should not be interrupted. When the candidate has been speaking for 2 minutes, the interviewer finds a logical way (at the end of a sentence or thought) to stop the candidate and moves on to the questions provided by the script. If the candidate finishes the monologue earlier than 2 minutes, the interviewer asks if the candidate has said everything he/she wanted to say and proceeds with the questions provided. All the questions should be asked in the same order they appear in the script. If the candidate has already answered any of the questions in their monologue, they should not be asked again. The monologue (with preparation time) and the discussion together should last for 8–9 minutes.

STAGE 3:

The interviewer follows the **Script for Stage 3.** The interviewer gives the candidate the *student card* with the task and instructs him/her to read it. Preparation time is 1 minute. When 1 minute has elapsed, the interviewer prompts the candidate to start. The interviewer uses the information in the script to answer candidate's questions. The interviewer should avoid long answers and give only the information the candidate has asked for. When the candidate does not conclude the role-play as required, the interviewer ends it by asking if the candidate has said all he/she wanted to say. This stage should take 4–5 minutes.

CLOSING: The interviewer rounds up the interview by thanking the candidate and stating that the interview is over. They should avoid evaluative comments which might give the candidate an idea about their examination results. The interviewer switches off the cassette recorder. The interviewer asks the candidate to sign the attendance form.

RECORDING: When interviews are recorded, the candidate's code numbers and the date of the recording should be written clearly on the cover of the cassette.

AFTER THE END OF THE EXAM: The assessor fills in the **ASSESSMENT FORMS (*PROTOKOLLID*)** as required, both the interviewer and the assessor have to sign them and the originals should be sent to the Examinations and Qualifications Centre. Copies of assessment forms can be kept at school. The assessor and interviewer pack up the examination materials, store them in the safe and destroy candidates' notes.

PHOTOCOPYING OF THE EXAM MATERIALS IS NOT ALLOWED!

Exception – the marking scale can be copied if the assessor wants to write their comments on the scale while assessing.

INTERVIEWERS' AND ASSESSORS' PROCEDURES 2008

Interviewer's procedures

Before the interviews

• arrive 45 minutes before to familiarize yourself with the examination materials for that day
• arrange the examination room to make it supportive
• test the cassette recorder to see if it works properly
• make sure there is paper and pens for the candidates to take notes with
• make sure there is a clock in the room for you to keep time

NB! The clock should only be visible to you, ideally behind the candidate, where you can look at it without disturbing the candidate.

• place twelve student cards for Stage 2 face down on the table

Before starting Stage 1

• greet the candidate in a friendly way
• ask the external candidates if they are familiar with the procedure / explain if necessary
• ask if the candidate wants the interview to be recorded. When the answer is 'Yes', switch on the cassette recorder and state the candidate's code number. When the answer is 'No', ask if the candidate is aware that he/she can only appeal against the result of the speaking paper if the answer is recorded.
• when you record interviews, the candidate's code numbers and the date of the recording should be written clearly on the cover of the cassette

Stage 1

• see the **Script for Stage 1**
• introduce the assessor to the candidate
• follow the script, do not improvise or paraphrase

Stage 2

• see the **Script for Stage 2**
• follow the script
• ask the candidate to choose a monologue card
• give the candidate some time to read the task
• give the candidate 3 minutes to plan his/her monologue

NB! the cassette recorder should not be switched off during that time

• when 3 minutes have elapsed or the candidate is ready to start, ask the candidate to start
• allow the candidate 2 minutes of uninterrupted monologue time
• when the candidate has been speaking for 2 minutes, find a logical way (at the end of a sentence or thought) to stop the candidate in a natural and friendly manner
• when the candidate has spoken for less than 2 minutes and it is not clear if he/she has finished, ask '*Is that all you wanted to say?' or 'Was there something else you wanted to say?'*
• continue with the questions in the script in the same order as they appear (unless the

candidate has already answered any of them in his/her monologue, in which case skip the question)
• once the questions have been answered, signal the end of the task by *'Thank you. Let's move on to the next task.'*

Stage 3

• see the **Script for Stage 3**
• give the candidate the card
• give the candidate 1 minute to think about the task
• when 1 minute is up, prompt the candidate to start by *'Could you start the role-play now.'*
• use the information in the script to answer candidate's questions
• keep your answers short
• do not give more information than the candidate asks
• if the candidate does not finish the role play as required (does not signal the decision at the end), ask *'Is that all you wanted to say?'*
• when the candidate has finished the role play, say, *'Thank you. This is the end of the interview.'*
• switch off the cassette recorder

Before the candidate leaves the room

• tell the candidate when the scores will be announced
• ask the candidate to sign the attendance form
• collect the candidate's notes

After the exam

• together with the assessor pack examination materials and destroy candidates' notes

The interviewer should

• be a friendly and attentive listener
• be natural
• keep to the wording of the stages given in the scripts
• avoid evaluative comments (e.g. good, well done, that was excellent, that's not very good, is it?, that's not right, you have not said very much)
• move on to the next question if the candidate is not willing to answer a question because of some personal reason
• keep to the time set for each part of the interview

The interviewer should not

• interrupt the candidate's monologue
• impose his/her views
• talk too much / speak more than the candidate
• enter into lengthy discussions with the candidate
• correct mistakes
• show with his/her body language that there has been a mistake (if a mistake occurs, continue in a friendly way as if nothing has happened)
• fill in the pauses when the candidate is clearly looking for words or ideas

Assessor's procedures

Before the examination the assessor should

• arrive at school 45 minutes before the start of the interview to familiarise him/herself

with the materials for that day
• assist the interviewer arranging the examination room
• make sure all the necessary documentation is there (scripts, evaluation forms, marking scale)
• make sure there is paper for taking notes of candidates' performance

During the examination the assessor should

• sit so that they can clearly hear the candidate (but interfere with his/her presence as little as possible)
• be as inconspicuous as possible
• apart from greeting the candidate, not interact
• make sure to record candidate's code number in Stage 1
• make sure to record the number of the topic of the candidate in Stage 2
• use all the criteria in the marking scale to assess every candidate's performance during Stages 2 and 3
• check against the script that the candidate has completed the tasks
• make notes to evaluate candidate's performance
• decide on the mark of each candidate immediately after the candidate has finished his/her interview

NB! The assessor can and should remind the interviewer of the correct procedural behaviour should the need arise. This can only be done when the candidate is **not** in the room.

After the examination the assessor should

• fill in the assessment form (*protokoll*)
• sign the form with the interviewer
• store the form safely with the materials until the last day of the oral part of the examination
• help the interviewer to pack up the exam materials and destroy candidates' notes
• keep his/her notes of candidates' performance for reference if a need should arise.

APPENDIX 3. QUESTIONNAIRE

TALLINNA ÜLIKOOL

English Language Teacher/Examiner Questionnaire

This spring, a new format was introduced for testing speaking at the English language national examination. You were probably one of the people who implemented the new format either as an interviewer or an assessor.

In order to fine-tune the examination process, making it more user-friendly for both examiners and students, test developers would highly appreciate your input. Please help us by answering the questions below as precisely as you can.

Please tick the box that best corresponds to your point of view:

1– NOT at all true **2**– mostly NOT true **3**– hard to say **4**– mostly true **5**– absolutely true

No.	Problem	1	2	3	4	5
1.	I was clear about the exam procedure before the exam started.					
2.	Pre-exam training was sufficient.					
3.	The examiner training materials were helpful.					
4.	I would need more training in the exam procedures.					
5.	I am willing to take part in an additional training course.					
6.	I check the examination centre website frequently for new materials.					
7.	Having a script for the interview was helpful.					
8.	It was easy to keep to the wording.					
9.	The wording of the frames seemed artificial.					
10.	I changed the wording of the script.					
11.	I ignored the wording of the frames completely.					
12.	Keeping time required effort.					
13.	I kept to the required timeframe.					

14.	It was easy to stop the students.					
15.	I forgot about the time.					
16.	I let the students talk for as long as they wanted.					
17.	Students understood what they had to do.					
18.	Students asked you to clarify what they needed to do					
19.	Students asked you to explain words.					
20.	Students took notes.					
21.	Students were ready before the given preparation time.					
22.	Students required more time than they were given.					
23.	Students used all the 2 minutes for the monologue.					
24.	Students finished before 2 minutes were over.					
25.	Students wanted to talk more.					
26.	Monologue topics were easily understandable.					
27.	Students found it easy to express their opinion on the topics.					
28.	Monologue topics were age appropriate.					
29.	Monologue topics were gender appropriate.					
30.	The follow-up questions were helpful.					
31.	The follow-up questions were appropriate.					
32.	Role-play is a good task-type for the speaking exam.					
33.	The topics for the role-play are appropriate.					
34.	I use the exact wording, answering students' questions in role-play.					
35.	I try to change the answer depending on the question.					
36.	It is easy to use the marking scale for speaking.					
37.	I need more practice with the marking scale.					
38.	I get nervous when the interview is recorded.					
39.	I record student interviews in class.					
40.	I can choose the classroom for the speaking exam.					

Please add any other comments regarding

a) training and materials

……
……
……
……
……

b) exam procedure

……
……
……
……
……

c) marking scale

……
……
……
……
……

d) any other aspect of the speaking exam

……
……
……
……
……

Please enclose the following details about yourself.

Gender: female /male (please circle)

Teaching experience in upper-secondary school/gymnasium (please circle):

1–2 years

3–5 years

6–10 years

11–15 years

16–20 years

more than 20 years

Number of classes taught per week in gymnasium .....

Thank you!

APPENDIX 4. STATISTICAL ANALYSES. TABLES

1. CHI-SQUARE TESTS AND PHI

1. 1. Comparison of schools: presence of introduction.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,066[a]	1	,302		
Continuity Correction[b]	,548	1	,459		
Likelihood Ratio	1,071	1	,301		
Fisher's Exact Test				,385	,230
Linear-by-Linear Association	1,044	1	,307		
N of Valid Cases	47				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,77.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,151	,302
	Cramer's V	,151	,302
N of Valid Cases		47	

1. 2. Comparison of schools: presence of greeting.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7,567[a]	1	,006		
Continuity Correction[b]	6,003	1	,014		
Likelihood Ratio	7,770	1	,005		
Fisher's Exact Test				,008	,007
Linear-by-Linear Association	7,406	1	,006		
N of Valid Cases	47				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 8,43.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,401	,006
	Cramer's V	,401	,006
N of Valid Cases		47	

1. 3. Comparison of schools: asking about well-being.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2,446[a]	1	,118		
Continuity Correction[b]	1,549	1	,213		
Likelihood Ratio	2,464	1	,117		
Fisher's Exact Test				,201	,107
Linear-by-Linear Association	2,394	1	,122		
N of Valid Cases	47				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,55.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,228	,118
	Cramer's V	,228	,118
N of Valid Cases		47	

1. 4. Gender comparison: Back-channeling.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13,021[a]	1	,000		
Continuity Correction[b]	10,547	1	,001		
Likelihood Ratio	13,746	1	,000		
Fisher's Exact Test				,001	,001
Linear-by-Linear Association	12,760	1	,000		
N of Valid Cases	50				

a. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 4,00.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,510	,000
	Cramer's V	,510	,000
N of Valid Cases		50	

1. 5. Comparison of school-types: Back-channeling.

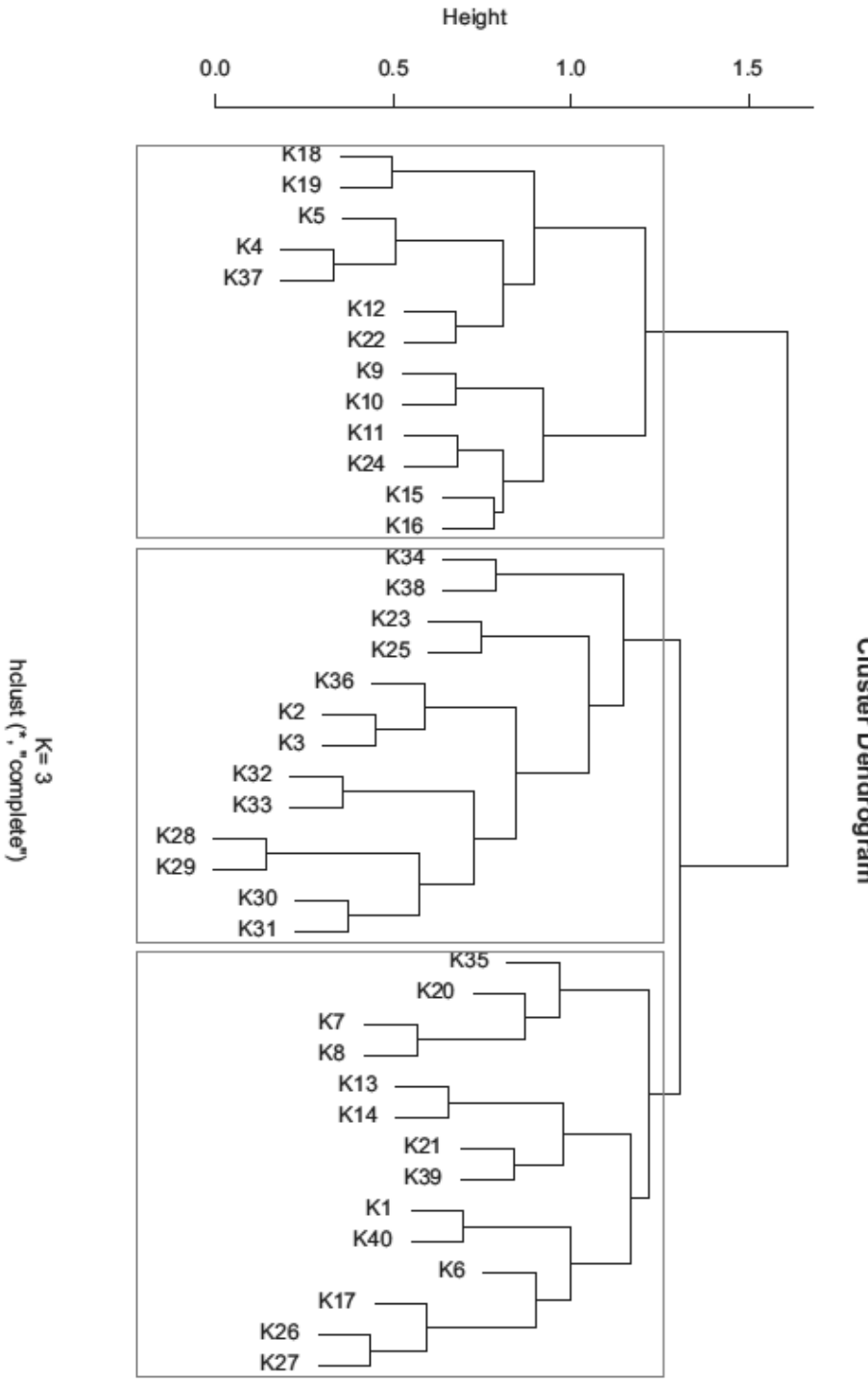
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,333[a]	1	,248		
Continuity Correction[b]	,750	1	,386		
Likelihood Ratio	1,340	1	,247		
Fisher's Exact Test				,387	,193
Linear-by-Linear Association	1,307	1	,253		
N of Valid Cases	50				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,00.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,163	,248
	Cramer's V	,163	,248
N of Valid Cases		50	


Height
0.0
0.5
1.0
1.5
K18
K19
K5
K4
K37
K12
K22
K9
K10
K11
K24
K15
K16
K34
K38
K23
K25
K36
K2
K3
K32
K33
K28
K29
K30
K31
K35
K20
K7
K8
K13
K14
K21
K39
K1
K40
K6
K17
K26
K27
K= 3
hclust (*, "complete")
Cluster Dendrogram

3. SPEARMAN'S RHO

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	K14	K15	K16	K17	K18	K19	K20	K21
K1	1,00																				
K2	,241*	1,00																			
K3	0,18	,550**	1,00																		
K4	-0,13	-0,19	-0,14	1,00																	
K5	0,12	-0,11	-0,13	,590**	1,00																
K6	0,10	-0,01	0,08	-0,15	-0,07	1,00															
K7	0,15	0,03	,251*	-0,01	-0,13	0,05	1,00														
K8	0,16	0,10	,372**	-0,15	-0,09	0,20	,435**	1,00													
K9	-,306**	-0,02	-,247*	0,10	-0,02	-0,13	-,226*	-,425**	1,00												
K10	-0,14	-0,07	-0,12	0,18	0,19	-0,03	0,04	-0,20	,327**	1,00											
K11	-0,20	0,11	-0,07	0,06	0,07	-0,05	-0,06	-0,22	0,08	,297**	1,00										
K12	-0,07	-,237*	-0,16	,288*	0,19	-0,23	0,09	-,242*	,259*	0,09	0,09	1,00									
K13	-0,02	-0,08	0,06	-0,03	-0,06	0,08	-0,06	-0,09	0,12	-0,17	-0,09	-0,02	1,00								
K14	-0,07	-0,02	0,04	-0,10	-0,03	0,13	-0,17	0,00	0,10	0,00	-0,06	-0,19	,345**	1,00							
K15	-0,21	-0,06	-,252*	0,17	0,05	-,231*	0,08	-0,12	0,13	0,11	0,19	0,20	-0,22	-,226*	1,00						
K16	-0,06	-0,21	-0,14	0,20	0,06	0,00	0,07	-0,08	0,12	0,18	,229*	0,07	0,05	-,255*	0,21	1,00					
K17	0,08	0,18	0,08	-,334**	-,318**	0,16	-0,13	0,08	-,266*	-0,18	-0,14	-0,18	0,07	,233*	-,297**	-,258*	1,00				
K18	-,248*	-0,21	-0,18	,246*	,274*	0,03	0,01	-0,18	0,10	,236*	0,07	0,17	-0,06	-0,11	0,11	0,17	-,432**	1,00			
K19	-0,14	-,326**	-,250*	0,17	,274*	-0,06	0,13	-0,04	0,00	0,13	0,08	,222*	0,01	-0,11	0,13	0,10	-,414**	,503**	1,00		

K20	0,19	0,13	0,15	0,19	0,11	0,03	0,14	0,13	-0,05	-0,11	-0,13	0,13	0,02	-0,16	-0,20	-0,01	0,06	0,08	-0,14	1,00	
K21	-0,09	-0,08	0,15	0,00	-0,16	0,13	0,09	0,04	-0,11	0,05	0,19	-,250*	0,10	,316**	-0,08	0,05	0,04	0,11	-0,08	-0,13	1,00
K22	-0,05	-0,16	-0,10	,241*	0,20	0,01	-0,01	-0,20	,316**	,259*	-0,21	,322**	-0,08	-0,08	,224*	0,00	-,405**	0,11	0,10	0,13	-,347**
K23	0,03	0,13	0,21	-0,08	-0,12	-0,05	0,05	0,12	-0,04	-0,06	-,292**	-0,07	0,18	0,16	-0,12	-,253*	0,11	-0,18	-,294**	,234*	-0,10
K24	-0,14	-0,07	-0,10	0,08	0,01	-0,09	-0,01	-0,20	0,11	,290**	,321**	0,11	-,226*	-0,07	0,21	0,19	0,04	-0,03	0,12	-,268*	0,17
K25	-0,02	0,10	0,02	0,04	-0,19	0,05	0,01	-,232*	,237*	0,06	-0,07	0,12	0,08	-0,12	0,06	0,09	0,03	-0,11	-0,19	0,02	-0,08
K26	,249*	0,16	0,19	-0,16	-0,02	0,10	0,03	,261*	-,363**	-,255*	-0,04	-,259*	-0,03	,313**	-,226*	-,309**	,473**	-,361**	-,263*	0,02	0,18
K27	,344**	,419**	,348**	-,261*	-0,14	0,13	-0,12	0,16	-,293**	-,245*	-0,15	-,413**	0,01	,241*	-,234*	-,405**	,404**	-,346**	-,404**	0,02	0,17
K28	,250*	0,19	,227*	-,245*	-0,12	0,04	0,14	0,21	-,369**	-0,17	-0,02	-0,11	0,07	0,13	-0,22	-0,10	,347**	-,333**	-0,15	-0,12	0,05
K29	,265*	0,21	,243*	-0,19	-0,10	0,01	0,12	,267*	-,413**	-,250*	-0,11	-0,21	0,05	0,02	-,241*	-0,06	,290*	-,280*	-0,10	-0,01	0,07
K30	0,18	,392**	,311**	-0,13	-0,12	0,06	0,13	0,21	-0,19	-0,22	-0,14	0,00	-0,04	-0,02	-0,17	-0,15	,322**	-,287**	-,309**	0,12	-0,16
K31	,284*	,266*	,404**	-0,17	-0,13	0,06	,306**	0,21	-,321**	-0,08	0,11	-0,05	-0,01	-0,06	-0,06	0,05	,282*	-,250*	-0,16	0,07	0,11
K32	,250*	0,16	,272*	-0,02	0,10	0,12	,312**	0,20	-0,13	-0,04	0,03	-0,11	0,07	-0,06	-0,19	0,07	0,04	-0,15	-0,02	0,17	-0,01
K33	0,11	0,15	,336**	-0,04	-0,10	0,13	,290**	,314**	-,236*	-0,06	0,07	-0,20	0,11	0,07	-0,13	0,09	,270*	-,297**	-0,18	0,03	0,11
K34	0,22	0,00	0,08	-0,04	-0,09	0,04	,303**	0,10	-0,05	-,295**	-0,13	0,20	-0,05	-0,13	0,05	0,15	0,10	-,278*	-0,03	-0,07	0,04
K35	0,01	0,11	0,20	-0,20	-0,12	0,10	0,03	0,11	-0,01	0,17	-0,13	0,00	-0,22	-0,14	0,03	0,08	0,02	-0,04	0,13	0,04	-0,10
K36	0,21	,472**	,412**	-,294**	-0,19	,245*	0,12	,257*	-0,07	-0,18	-0,06	-,304**	0,13	0,09	-0,19	-,258*	0,12	-0,16	-0,15	0,09	0,04
K37	0,04	-0,08	-0,01	,668**	,494**	-,256*	0,04	-0,09	-0,04	0,01	-0,03	,365**	-0,19	-0,14	0,17	-0,10	-,313**	0,20	,260*	0,16	-0,17
K38	-0,21	-0,05	-0,05	0,15	0,02	-0,07	0,11	0,03	0,09	0,01	0,07	,257*	-,305**	-0,12	0,02	0,15	-0,04	-0,09	0,09	-0,01	-,229*
K39	0,07	0,06	0,07	0,06	-0,15	-0,05	-0,02	-0,02	0,12	-0,09	-0,09	-0,04	0,04	0,02	-0,10	-0,12	-0,03	0,00	-0,02	-0,05	0,16
K40	,303**	0,04	0,01	-0,03	0,18	0,07	-0,10	0,05	0,03	-0,05	0,00	0,09	0,15	-0,03	-0,08	-0,05	0,00	0,01	0,07	0,04	-0,17
TE	0,05	0,04	0,02	-0,18	-,280*	-,236*	0,08	0,04	-0,07	-0,06	0,02	-0,10	0,08	-0,13	-0,04	-0,02	0,10	-0,09	-0,08	-0,07	-0,01

	K22	K23	K24	K25	K26	K27	K28	K29	K30	K31	K32	K33	K34	K35	K36	K37	K38	K39	K40	TE
K22	1,00																			
K23	0,14	1,00																		
K24	-0,05	-,612**	1,00																	
K25	,230*	,254*	-0,11	1,00																
K26	-,314**	0,09	-0,09	-,275*	1,00															
K27	-,278*	0,15	-0,16	0,07	,565**	1,00														
K28	-,287*	0,04	0,03	0,00	,433**	,500**	1,00													
K29	-,294**	0,03	0,02	0,04	,383**	,541**	,858**	1,00												
K30	-0,17	0,19	-0,11	0,13	,277*	,299**	,503**	,425**	1,00											
K31	-,290**	0,05	0,07	-0,05	,247*	,264*	,469**	,448**	,629**	1,00										
K32	-0,14	0,02	-0,06	-0,04	0,11	0,15	,272*	,313**	,337**	,456**	1,00									
K33	-,343**	-0,01	0,03	-0,04	0,16	0,15	,346**	,405**	,391**	,589**	,642**	1,00								
K34	-0,10	-0,12	0,16	0,10	0,15	-0,02	0,09	0,10	0,16	0,15	0,15	0,21	1,00							
K35	0,16	-0,10	0,11	0,06	-0,19	0,00	-0,02	0,11	0,03	0,15	-0,04	0,03	-0,18	1,00						
K36	-0,22	,253*	-0,19	0,12	0,08	,255*	0,20	,226*	,414**	,240*	,290**	,345**	0,00	-0,01	1,00					
K37	,307**	-0,04	0,07	-0,09	-0,02	-0,14	-0,19	-0,15	-0,17	-0,15	-0,06	-0,15	0,11	-0,14	-,392**	1,00				
K38	0,09	-0,02	0,15	0,01	-0,08	-,290**	-0,11	-0,15	0,13	0,00	-0,09	0,01	0,21	0,16	-0,03	0,15	1,00			
K39	0,09	0,21	-0,15	0,13	-0,06	0,12	-0,13	-0,14	-0,18	-0,18	-0,13	-0,13	0,02	-0,11	0,09	-0,01	-0,05	1,00		
K40	-0,05	-0,01	-0,18	0,11	0,02	0,11	0,16	0,05	0,01	-0,04	0,01	-0,17	-0,10	-0,08	0,07	-0,02	-0,08	-0,01	1,00	
TE	-0,09	,266*	-,234*	0,10	-0,21	-0,05	-0,10	-0,08	0,09	0,07	0,03	0,02	-0,08	0,06	0,13	-0,20	0,06	0,22	0,07	1,00

* Correlation is significant at the 0.05 level (2-tailed)
** Correlation is significant at the 0.01 level (2-tailed)

4. T-TESTS:

4. 1. Overall interview duration by school-type (Estonian – Russian).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Overall Time	Equal variances assumed	,087	,769	,516	44	,609	,22968	,44554	-,66825	1,12762
	Equal variances not assumed			,518	43,415	,607	,22968	,44308	-,66362	1,12299

4. 2. Interview duration by gender.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Overall Time	Equal variances assumed	,529	,471	-2,302	44	,026	-1,22027	,53011	-2,28863	-,15191
	Equal variances not assumed			-2,399	12,846	,032	-1,22027	,50856	-2,32030	-,12024

4. 3. Interview duration by gender. Russian schools.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Overall-Time	Equal variances assumed	,310	,583	-1,171	23	,253	-,80093	,68383	-2,21554	,61369
	Equal variances not assumed			-1,248	12,568	,235	-,80093	,64202	-2,19279	,59094

4. 4. Interview duration by gender. Estonian schools.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Overall Time	Equal variances assumed	3,764	,067	-2,472	19	,023	-2,38333	,96425	-4,40153	-,36513
	Equal variances not assumed			-4,869	2,600	,023	-2,38333	,48944	-4,08557	-,68110

ELULOOKIRJELDUS

1. **Ees- ja perekonnanimi:** Ene Alas
2. **Sünniaeg ja koht**: 19. mai 1959, Tallinn, Eesti
3. **Kodakondsus**: Eesti
4. **Kontaktaadress ja telefon**: enealas@tlu.ee

5. **Haridus**: kõrgem, magister, inglise kui võõrkeele õpetamine, 1994,
   1993–1994 Magistrikursused, Reading University, magistrikraad, Suurbritannia 1994,
   1977–1982 TPed I inglise js saksa keele õpetaja, diplom, 1982.

6. **Läbitud täiendusõpe**:
   2010 Rahvusvahelise IELTS testija rekvalifikatsioonikursus, Tallinn.
   2008 Rahvusvahelise IELTS testija rekvalifikatsioonikursus, Tallinn.
   2007 Briti Nõukogu koolitusprogramm 'Continuing Professional Development'.
   2007 Briti Nõukogu koolitusprogramm 'Presentation and Training Skills'.
   2006 Briti Nõukogu ja REKK'i koolitusprogramm 'Assessment Across the Curriculum'.
   2006 Rahvuvahelise IELTS testija kvalifikatsioonikursused, Riia, Läti.
   2005 Individuaalne täiendõppekursus King's College London, Suurbritannia.
   2004 Express Publishing Opening Seminar, Tallinn.
   2003 British Council Database Course: OU and OED Database. Briti Nõukogu, Tallinn.
   2000–2001 doktorantide kursused Department of Applied Linguistics, Indiana University.
   1993–1994 Magistrikursused, University of Reading, Suurbritannia.
   1991 Briti Nõukogu kursused inglise keele kui võõrkeele õppejõududele, Colchester University, Suurbritannia.
   1988 Diplomikursus 'Certificate Course in the Language and Cultural Aspects of Teaching English at the University Level'. University of Guildford, Suurbritannia.

7. **Teenistuskäik:**
   Aug. 2002 – praeguseni – TLÜ GRKKI lektor,
   Aug. 2001 – aug.2002 Tallinna Euroülikooli lektor,
   Aug. 2000 – aug. 2001 Külalislektor, CEUS, Indiana Univeristy, USA,
   Aug. 1998 – aug. 2000 TPÜ inglise keele õppetooli lektor,
   Aug. 1996 – aug. 1998 Külalislektor, CEUS, Indiana University, USA,
   Aug. 1989 – aug. 1996; TPÜ inglise keele õppetooli lektor,
   1987–1989 TPÜ inglise keele õppetooli õpetaja,
   1984–1987 TPÜ inglise keele õppetooli tunnitasuline õppejõud,
   1982–1984 Tallinna 32. Keskkool, inglise keele õpetaja.

8. Kehtivad töösuhted:
TLÜ, põhikohaga õppejõud, 20. aug. 2009.
TLÜ eksamikeskus, rahvusvahelise kategooria IELTS eksamineerija.

9. Keelteoskus: eesti, inglise, vene, soome, saksa.

10. Viimase viie aasta tegevus:

1. GRKKI kolleegiumi liige alates 2007, õpetajakoolituse nõukogu liige alates 2007.
2. Uurimisgrandid: 2005 Erasmuse stipendium õppetööl osalemiseks King's College London'is, Suurbritannia.
3. 2009 Communication Skills Workshop: Tampere. Korralduskomitee Eesti-poolne liige.
4. Konverentsiettekanded:
   2009. TLÜ GRKKI konverents. Ettekanne: Interviewer Variability in Oral Proficiency Interviews.
   2008. EATE Pärnu konverents. Ettekanne: A Decade of National Examinations in English in Estonia.
   2008. EATE Tartu konverents. Ettekanne: National Examination Results 2008.
   2007. TU Narva kolledž. Konverents 'The secrets of ESP/EAP'. Ettekanne: Working towards style in academic writing.
   2007. EATE Pärnu konverents. Ettekanne: Changes in the Oral Part of the National Examination 2008.
   2006. CSW 21st Workshop. Ettekanne: Bridging Theory and Practice.
   2005. 2nd ALTE International Conference: Berlin. Ettekanne: Assessing Academic Writing.
   2004. Communication Skills 19th Workshop. Ettekanne: Writing for the Academia. Nastola, Soome.
5. Välissidemed: King's College London, UK; Communication Skills Workshop, Soome; Reading University, UK; Indiana University USA.
6. Koostatud õppevahendid:
   Alas, Ene (2008). Guidelines for the Oral Part of the National Examination 2008. Riikliku Eksami ja Kvalifikatsioonikeskuse koduleheküIg.
   Alas, Ene, Roosmaa, Ester (2008). Interviewers' and Assessors' Procedures for the Nationa Examination 2008. Riikliku Eksami ja Kvalifikatsioonikeskuse koduleheküIg.
   Alas, Ene (2008). National Examination Marking Scale for Speaking. Riikliku Eksami ja Kvalifikatsioonikeskuse koduleheküIg.
   Alas, Ene (2008). Oral Examination Introductory Stage Specimen. Riikliku Eksami ja Kvalifikatsioonikeskuse koduleheküIg.

Alas, Ene (2007). A Marking Scale for Letters, A Marking Scale for Essays and Reports. National Examination 2007. Riikliku Eksami ja Kvalifikatsioonikeskuse koduleheküg.

7. Juhendatud/retsenseeritud magistritööd:
   - 2009 Irina Biba. Designing a Final Achievement Test for Young Learners. Juhendaja.
   - 2009 Kristiina Toots. Teaching Vocabulary to Young Adults on the Basis of "New Headway" Series. Juhendaja.
   - 2009 Jelena Šmidt. Neuro-Linguistic Programming in English Language Teaching. Juhendaja.
   - 2008 Jekaterina Wilde. Teaching Critical Reading Skills at Secondary School Level. Retsensent.
   - 2008 Margit Kirss. Formative Aspects of Assessing Listening Comprehension of Young Learners. Juhendaja.
   - 2008 Olga Gabovits-Behhalova Task-Based Language Teaching Approach for Primary Level: from Theory to Practice. Retsensent.
   - 2007 Evelyn Soidla. Materials Development for ESP Courses: The Case of Rescue Speciality in the Public Service Academy. Juhendaja.
   - 2007 Janika Johanna Marley. Discourse Analysis and Language Teaching: An Analysis of Vocabulary Diversity in the Transcripts of 'The Bold and the Beautiful' and in Practical Applications to Language Teaching. Juhendaja.
   - 2007 Anu Joon. Using Total Immersion method in Young Learner Instruction. Retsensent.
   - 2007 Janne Rahula. Music and Drama as Means of Learning. Retsensent.
   - 2006 Dimitri Leontjev. Testing receptive Skills in a foreign Language. Juhendaja.
   - 2004 Mari Martma. Foreign Language Acquisition at Pre-School Level: From Theory to Practice. Retsensent.

CURRICULUM VITAE

1. **Name:** Ene Alas
2. **Date of birth:** May 19, 1959, Tallinn, Estonia
3. **Citizenship:** Estonian
4. **Contact:** enealas@tlu.ee

5. **Education:** MA in TEFL, 1994,
   1993–1994 Master's Courses, Reading University, UK. MA degree, 1994,
   1977–1982 Tallinn Teacher Training Institute, diploma of English and German teacher 1982.

6. **In-service training:**
   2010 Requalification course for the international IELTS tester, Tallinn.
   2008 Requalification course for the international IELTS tester, Tallinn.
   2007 British Council training programme 'Continuing Professional Development'.
   2007 British Council training programme 'Presentation and Training Skills'
   2006 British Council and NEQC training programme 'Assessment Across the Curriculum'.
   2006 Qualification Course for the international IELTS tester Riga, Latvia.
   2005 Erasmus individual research course King's College London, UK.
   2004 Express Publishing Opening Seminar, Tallinn.
   2003 British Council Database Course: OU and OED Database. British Council, Tallinn.
   2000–2001 Doctoral courses Department of Applied Linguistics, Indiana University, USA.
   1993–1994 M.A. Courses in TEFL University of Reading, United Kingdom.
   1991 British Council Certificate Course in Specialized Teaching of English as a Foreign Language. University of Colchester, United Kingdom.
   1988 Certificate Course in the Language and Cultural Aspects of Teaching English at the University Level. University of Guildford, United Kingdom.

7. **Career:**
   Aug. 2002 – present time – Tallinn University lecturer of English and testing,
   Aug. 2001 – aug.2002 Tallinn Eurouniversity lecturer,
   Aug. 2000 – aug. 2001 visiting lecturer, CEUS, Indiana University, USA,
   Aug. 1998 – aug. 2000 TPU lecturer,
   Aug. 1996 – aug. 1998 visiting lecturer, CEUS, Indiana University, USA,
   Aug. 1989 – aug. 1996; TPU lecturer,
   1987–1989 TPU teacher,
   1984–1987 TPU part time teacher,
   1982–1984 Tallinna Secondary School No. 32. English teacher.

8. Current employers:
Tallinn University, Institute of Germanic and Romance Languages and Cultures, lecturer of English and testing,
Tallinn University, international IELTS tester.

9. Languages: Estonian, English, Russian, Finnish, German.

10. Additional academic activities (past five years):

1. Member of IGRLC board and teacher education council as of 2007.
2. Research grants: 2005 Erasmus grant for King's College London, UK.
3. 2009 Communication Skills Workshop: Tampere. Member of Organising Committee.
4. Conference presentations:
   2009. Tallinn University: Interviewer Variability in Oral Proficiency Interviews.
   2008. EATE Pärnu: A Decade of National Examinations in English in Estonia.
   2008. EATE Tartu: National Examination Results 2008.
   2007. TU Narva college conference 'The secrets of ESP/ EAP': Working towards style in academic writing.
   2007. EATE Pärnu: Changes in the Oral Part of the National Examination 2008.
   2006. CSW 21st Workshop: Bridging Theory and Practice.
   2005. 2nd ALTE International Conference: Berlin: Assessing Academic Writing.
   2004. Communication Skills 19th Workshop: Writing for the Academia. Nastola, Finland.
5. International Contacts: King's College London, UK; Communication Skills Workshop, Finland; Reading University, UK; Indiana University USA.
6. Materials Developed:
   Alas, Ene (2008). Guidelines for the Oral Part of the National Examination 2008. www.ekk.edu.ee.
   Alas, Ene, Roosmaa, Ester (2008). Interviewers' and Assessors' Procedures for the Nationa Examination 2008. www.ekk.edu.ee.
   Alas, Ene (2008). National Examination Marking Scale for Speaking. www.ekk.edu.ee.
   Alas, Ene (2008). Oral Examination Introductory Stage Specimen. www.ekk.edu.ee.
   Alas, Ene (2007). A Marking Scale for Letters, A Marking Scale for Essays and Reports. National Examination 2007. www.ekk.edu.ee.

7. Master's Theses supervised and reviewed:
   2009 Irina Biba. Designing a Final Achievement Test for Young Learners. Supervisor.
   2009 Kristiina Toots. Teaching Vocabulary to Young Adults on the Basis of "New Headway" Series. Supervisor.
   2009 Jelena Šmidt. Neuro-Linguistic Programming in English Language Teaching. Supervisor.
   2008 Jekaterina Wilde. Teaching Critical Reading Skills at Secondary School Level. Reviewer.
   2008 Margit Kirss. Formative Aspects of Assessing Listening Comprehension of Young Learners. Supervisor.
   2008 Olga Gabovits-Behhalova Task-Based Language Teaching Approach for Primary Level: from Theory to Practice. Reviewer.
   2007 Evelyn Soidla. Materials Development for ESP courses: The Case of Rexcue Speciality in the Public Service Academy. Supervisor.
   2007 Janika Johanna Marley. Discourse Analysis and Language Teaching: An Analysis of Vocabulary Diversity in the Transcripts of 'The Bold and the Beautiful' and in Practical Applications to Language Teaching. Supervisor.
   2007 Anu Joon. Using Total Immersion method in Young Learner Instruction. Reviewer.
   2007 Janne Rahula. Music and Drama as Means of Learning. Reviewer.
   2006 Dimitri Leontjev. Testing receptive Skills in a foreign Language. Supervisor.
   2004 Mari Martma. Foreign Language Acquisition at Pre-School Level: From Theory to Practice. Reviewer.

TALLINN UNIVERSITY
DISSERTATIONS ON HUMANITIES.

TALLINNA ÜLIKOOL
HUMANITAARTEADUSTE DISSERTATSIOONID.

1. СЕРГЕЙ ДОЦЕНКО. *Проблемы поэтики А. М. Ремизова. Автобиографизм как конструктивный принцип творчества*. Таллинн: Изд-во ТПУ, 2000. 162 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам, 1. ISSN 1406-4391. ISBN 9985-58-135-0.

2. MART KIVIMÄE. *Ajaloomõtlemise kolm strateegiat ja nende dialoogisuhted minevikuga (lisades tõlgitud R. Koselleck, J. Rüsen, E. Nolte). Historismi muutumise, arendamise, ületamise probleemid.* Tallinn: TPÜ kirjastus, 2000. 201 lk. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid, 2. ISSN 1406–4391. ISBN 9985-58-164-4.

3. НАТАЛЬЯ НЕЧУНАЕВА. *Минея как тип славяно–греческого средневекового текста.* Таллинн: Изд-во ТПУ, 2000. 177 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам, 3. ISSN 1406-4391. ISBN 9985-58-125-3.

4. ОЛЕГ КОСТАНДИ. *Раннее творчество В. Каверина как литературный и культурный феномен.* Таллин: Изд-во ТПУ, 2001. 142 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам, 4. ISSN 1406–4391. ISBN 9985-58-180-6.

5. LAURI LINDSTRÖM. *Album Academicum Universitatis Tartuensis 1918–1944. Rahvus, sugu, sünnikoht ja keskhariduse omandamise koht üliõpilaskonna kujunemist ja kõrghariduse omandamist mõjutavate teguritena*. Tallinn: TPU Press, 2001. 92 p. Tallinn Pedagogical University. Dissertations on Humanities Sciences, 5. ISSN 1406-4391. ISBN 9985-58-190-3.

6. АУРИКА МЕЙМРЕ. *Русские литераторы-эмигранты в Эстонии 1918–1940. На материале периодической печати.* Таллин: Изд-во ТПУ, 2001. 165 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам, 6. ISSN 1406-4391. ISBN 9985-58-205-5.

7. AIVAR JÜRGENSON. *Siberi eestlaste territoriaalsus ja identiteet.* Tallinn: TPÜ kirjastus, 2002. 312 lk. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid, 7. ISSN 1406–4391. ISBN 9985-58-239-X.

8. DAVID VSEVIOV. *Kirde-Eesti urbaanse anomaalia kujunemine ning struktuur pärast Teist maailmasõda* Tallinn: TPÜ kirjastus, 2002. 104 lk. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid, 8. ISSN 1406-4391. ISBN 9985-58-242-X.

9. ROMAN KALLAS. *Eesti kirjanduse õpetamise traditsioon XX sajandi vene õppekeelega koolis.* Tallinn: TPÜ kirjastus, 2003. 68 lk. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid, 9. ISSN 1406–4391. ISBN 9985-58-256-X.

10. KRISTA KERGE. *Keele variatiivsus ja mine-tuletus allkeelte süntaktilise keerukuse tegurina.* Tallinn: TPÜ kirjastus, 2003. 246 lk. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid, 10. ISSN 1406-4391. ISBN 9985-58-265-9.

11. АННА ГУБЕРГРИЦ. *Русская драматургия для детей как элемент субкультуры: 1920–1930-е годы.* Таллинн: Изд-во ТПУ, 2004. 168 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам, 11. ISSN 1406–4391. ISBN 9985-58-302-7.

12. VAHUR MÄGI. *Inseneriühendused Eesti riigi ülesehituses ja kultuuriprotsessis (1918–1940).* Tallinn: TPÜ kirjastus, 2004. 146 lk. Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid, 12. ISSN 1406-4391. ISBN 9985-58-344-2.

13. HEIKKI OLAVI KALLIO. *Suomen ja Viron tiedesuhteet erityisesti Viron miehitysaikana vuosina 1940–1991.* Tallinn: Tallinnan Pedagogisen Yliopiston kustantamo, 2004. 243 lk. Tallinnan Pedagogisen Yliopiston. Humanististen tieteiden väitöskirjat, 13. ISSN 1406-4391. ISBN 9985-58-350-7.

14. ÜLLE RANNUT. *Keelekeskkonna mõju vene õpilaste eesti keele omandamisele ja integratsioonile Eestis.* Tallinn: TLÜ kirjastus, 2005. 215 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid, 14. ISSN 1406-4391. ISBN 9985-58-394-9.

15. MERLE JUNG. *Sprachspielerische Texte als Impulse für schriftliche Textproduktion im Bereich Deutsch als Fremdsprache.* Tallinn: Verlag der Universität Tallinn, 2006. 186 S. Universität Tallinn. Dissertationen in den Geisteswissenschaften, 15. ISSN 1406-4391. ISBN 9985-58-409-0

16. ANDRES ADAMSON. *Hertsog Magnus von Holmsteini roll Läänemere-ruumis Liivi sõja perioodil.* Tallinn: TLÜ kirjastus, 2005. 156 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid, 16. ISSN 1736-3624. ISBN 9985-58-427-9.

17. АИДА ХАЧАТУРЯН. *Роман В.С. Маканина «Андеграунд, или Герой нашего времени»: Homo urbanis в поле «усреднения»*. Таллинн: Изд-во ТПУ, 2006. 146 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам, 17. ISSN 1736–3624. ISBN-10 9985-58-435-X. ISBN-13 987-9985-58-435-4.

18. JULIA TOFANTŠUK. *Construction of Identity In The Fiction of Contemporary British Women Writers (Jeanette Winterson, Meera Syal, and Eva Figes).* Tallinn: Tallinn University Press, 2001. 160 p. Tallinn University. Dissertations on Humanities Sciences, 18. ISSN 1736-3624. ISBN 978-9985-58-479-8.

19. REILI ARGUS. *Eesti keele muutemorfoloogia omandamine.* Tallinn: TLÜ kirjastus, 2007. 242 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid, 19. ISSN 1736-3624. ISBN 978-9985-58-543-6.

20. ÕNNE KEPP. *Identiteedi suundumusi Eesti luules.* Tallinn: TLÜ kirjastus, 2008. 222 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid, 20. ISSN 1736-3624. ISBN 978-9985-58-559-7.

21. ANNELI KÕVAMEES. *Itaalia eesti reisikirjades: Karl Ristikivi „Itaalia capriccio" ja Aimée Beekmani „Plastmassist südamega madonna".* Tallinn: TLÜ kirjastus, 2008. 141 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid, 21. ISSN 1736-3624. ISBN 978-9985-58-574-0.

ILMUNUD VEEBIVÄLJAANDENA
http://www.tlulib.ee/?LangID=1&CatID=504

1. ИННА АДАМСОН. *Модальный смысл дезидеративности: от семантической зоны к семантической типологии высказываний (на материале русского языка)*. Таллинн: Изд-во ТЛУ, 2006. 131 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам. ISSN 1736-5031. ISBN 978-9985-58-455-2.

2. MARIS SAAGPAKK. *Deutschbaltische Autobiographien als Dokumente des zeit- und selbstempfindens: vom ende des 19. Jh. Bis zur umsiedlung 1939.* Tallinn: Verlag der Universität Tallinn, 2006. 163 S. Universität Tallinn. Dissertationen in den Geisteswissenschaften. ISSN 1736-5031. ISBN 978-9985-58-469-9.

3. JANIS EŠOTS. *Mullā Sadrā's Teaching on Wujūd: A Synthesis of Mysticism and Philosophy.* Tallinn: Tallinn University Press, 2007. 150 p. Tallinn University. Dissertations on Humanities Sciences. ISSN 1736-5031. ISBN 978-9985-58-492-7.

4. ГРИГОРИЙ УТГОФ. *Проблема синтактического темпа.* Таллинн: Изд-во ТЛУ, 2007. 145 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам. ISSN 1736-5031. ISBN 978-9985-58-507-8.

5. ДИМИТРИЙ МИРОНОВ. *Глагольность в сфере имен: к проблеме семантического описания девербативов* (на материале русского языка). Изд-во ТЛУ, 2008. 98 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам. ISSN 1736-5031. ISBN 978-9985-58-563-4

6. INNA PÕLTSAM-JÜRJO. Liivimaa väikelinn varase uusaja lävel. Uurimus Uus-Pärnu ajaloost 16. sajandi esimesel poolel. Tallinn: TLÜ kirjastus, 2008. 257 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid. ISSN 1736-5031. ISBN 978-9985-58-570-2.

7. TIIT LAUK. *Džäss Eestis 1918–1945*. Tallinn: TLÜ kirjastus, 2008. 207 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid. ISSN 1736-5031. ISBN 978-9985-58-594-8.

8. ANDRES ADAMSON. *Hertsog Magnus ja tema "Liivimaa kuningriik"*. Tallinn: TLÜ kirjastus, 2009. 173 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid. ISSN 1736-5031. ISBN 978-9985-58-615-0.

9. ОЛЕСЯ ЛАГАШИНА. *Марк Алданов и Лев Толстой: к проблеме рецепции*. Таллинн: Изд-во ТЛУ, 2009. 151 стр. Таллиннский педагогический университет. Диссертации по гуманитарным наукам. ISSN 1736-5031. ISBN 978-9985-58-654-9.

10. MARGIT LANGEMETS. *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras*. Tallinn: TLÜ kirjastus, 2009. 259 lk. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid. ISSN 1736-5031. ISBN 978-9985-58-651-8.

DISSERTATSIOONINA KAITSTUD MONOGRAAFIAD
(ilmunud iseseisva väljaandena)

1. ANNE VALMAS. *Eestlaste kirjastustegevus välismaal 1944–2000. I-II*. Tallinn: Tallinna Pedagoogikaülikooli kirjastus, 2003. 205, 397 lk. Tallinna Pedagoogikaülikool. ISBN 9985-58-284-5. ISBN 9985-58-285-3.
2. ANNE LANGE. *Ants Oras.* Monograafia. Tartu: Ilmamaa, 2004. 493 lk. ISBN 9985-77-163-X.
3. KATRI AASLAV-TEPANDI. *Eesti näitlejanna Erna Villmer*. Monograafia. Tallinn: Eesti Teatriliit, 2007. 495 lk. ISBN 78-9985-860-41-0.