



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



**An investigation of final language
assessment for pre-service teachers of
English in the Russian educational
context: a case study**

Volume I

Natalia Sokolova

Doctor of Philosophy

The University of Edinburgh

2015

Declaration of originality of submitted work

I hereby declare that this thesis, submitted in candidature for the degree of Doctor of Philosophy at the University of Edinburgh, and the research contained herein is of my own composition, except where explicitly stated in the text, and was not previously submitted for the award of any other degree or professional qualification at this or any other university.

Natalia Sokolova, MA, Cand (Kandidat Nauk) TESOL

Signed.....

Date.....

Abstract

This research explores the final assessment of language competence of future foreign language (FL) teachers (university graduates) in the Russian educational context. Foreign Language teacher training has always been an essential part of Russian education and its importance increased in the 1990s. Later however, with significant educational reforms at primary and secondary school level, teacher training became an area of least attention and interest from the Ministry of Education of Russia and local education authorities. This research is based on the belief that no school reforms are possible without investing in teachers and, therefore, in initial and in-service teacher education, with assessment being one of its key dimensions.

The study aims to describe optimal methods of assessing language competence of novice teachers of English as a FL in Russia. For this purpose, the following objectives have been achieved:

- a description of current notions of FL teacher language competence, based on analyses of previous theoretical and empirical research;
- design of exam evaluation tools – 3 questionnaires and an interview framework, and their use in data collection from various stakeholders in a Russian state pedagogical university;
- identification of strengths and weaknesses of the current Final language assessment;
- description of possible alternative options for the Final Language Examination and discussion of their impact on different stakeholders.

The research follows a mixed-methods design with both qualitative and quantitative data collected and discussed. The study involves various stakeholders at different levels and from different backgrounds – university students, Final Exam takers; Exam designers and administrators, and also teachers of English who provided their valuable vision of the current Final Language Examination and its possible alternatives. The data obtained through surveys and interviews allows for tentative conclusions on the current Language Examination's appropriacy and relevance, and provides ground for a multi-faceted analysis of the Exam's strong points and weaknesses, and for development of alternative assessment tasks.

The research concludes by viewing possible changes in the Exam as likely and less likely to happen in the near future, based on analysis of the Russian higher education context.

Lay summary

Name of student:	Natalia Sokolova	UUN	S0971345
University email:	s0971345@sms.ed.ac.uk		
Degree sought:	Doctor of Philosophy	No. of words in the main text of thesis:	93038
Title of thesis:	An investigation of final language assessment for pre-service teachers of English in the Russian educational context: a case study		

This research explores the Final Language Examination for future teachers of English as a foreign language (university graduates) in Russia. The research aims to analyze the existing Examination content, format and administration, and also to suggest possible alternatives of assessing language competence of novice language teachers. To do so, the following steps have been taken:

- review of the existing theoretical and empirical research on language teacher language competence, which allows for a taxonomy of skills, knowledge areas and attitudes that a modern foreign language teacher is expected to demonstrate;
- review of modern trends and tendencies in language teacher assessment and evaluation;
- design of the Exam evaluation tools – 3 questionnaires and an interview framework for various groups of stakeholders; data collection and analysis;
- identification of strengths and weaknesses of the current Final Language Assessment;
- description of possible alternative options for the Final Language Examination and discussion of their impact on different stakeholders – university students, school teachers of English, etc.

The study involves various methods of data collection and various stakeholders at different levels and from different backgrounds – university students, Final Exam takers; Exam designers and administrators, and also school teachers of English. The obtained data allows for tentative conclusions to be made on the strong and weak points of the current Examination, and suggestions for alternative Exam tasks and administration procedures.

Acknowledgements

I would like to thank my supervisors *Dr Aileen Irvine* and *Dr Rosemary Douglas* for their guidance, very useful feedback and helpful advice, for their expertise and interest in my research and my life, for encouragement and support. With so many new things to deal with in a foreign country and a foreign educational system, such help is precious.

My special thanks and deepest love go to my darling parents *Lyuba Sokolova(KMX)* and *Dr Hermann Sokolov* who never doubted I would achieve what I always wanted, and for whom I have always been ‘their Puss’ - whatever I do and wherever I am. Without their love, support and sacrifice, my PhD journey would be an absolutely different experience with different (and doubtful) outcomes. Words fail when I want to express my love and gratitude.

Heartiest thanks and lots of love to my dear friend and guardian angel *Barbara Udok* and her husband *Ekanem* for helping, supporting and encouraging me in every possible and impossible way, for believing in me and giving me strength and confidence, for being there for me 24/7 and for being such an important and wonderful part of my life. Thank you so much for your intelligence, your advice, kindness, humour and generosity; for every smile, tear and hug, for every secret kept.

To *Dr Maxim Shadurski* for fantastic times, for feedback and advice, for sharing his experience and tolerating me all these years – for jokes and desperation shared, unfinished sentences and silence understood, for encouragement and always being there.

A big thank you to my friends in Russia for keeping in touch, encouragement and good times spent together, for helping my parents and being there for them, for emails, Skype sessions, Facebook comments and all kinds of positive things that friendship brings: *Irina Belousova, Roman* and *Julia Martynyuk* and young *Michael; Tatiana* and *Polina Kodukova*.

My friend and colleague *Natalia Morozova* and pupils of Tula Comprehensive School #25; my former TESOL students, now my colleagues - *Maya Burik, Anna Matveyeva, Tatiana Vedeshkina, Anastasia Demidova, Maria Filina, Margarita Pervukhina, Elena Yakimova*.

My fellow doctorate researchers and good friends in Moray House School of Education – *Enid Quesada Alfaro, Patricia Cacho, Jane Spiteri, Wida Suhaili, Sumera Umrani, Yuchen Wang* for their support, advice and positive attitude that helped me a lot.

To the administrative staff and my colleagues at Brightcare at Home Ltd – for their understanding and patience, for positive environment and valuable experience which I hope to go on getting: *Tim Cocking, Elizabeth Hulbert, Agnieszka Gorak, Val Brown, David Fairweather, Dorothy-Ann Anderson, Jemima Vetha, Stephanie Harris*.

To my Brightcare clients *Dr. Michael Robson*, *Dr Elizabeth Swan*, *Allan McDonald*, *Catherine Sharp, Jenny Donaldson; Christine Pow, Ann Allan, Annie Stark*, and *Maureen Wishart*, the best ever co-ordinator at Lymedoch House, Edinburgh, for their understanding and support, for interest in my studies and opportunity ‘to do what is best for me’ and ‘to go home NOW and study’. Without all these, my research would have been a much more difficult path.

Table of contents

Abstract	
List of abbreviations used in this thesis	i
List of institutions, people and events referred to in the thesis	ii
The hierarchy of institutions in the system of Russian higher education and structure of a Russian pedagogical university	iii
List of tables	iv
List of figures and pictures	viii
Chapter 1. Introduction	1
Chapter 2. Context of the Research: Foreign Language Teacher Development in Russia	6
2.1. Changes in the Russian system of Foreign Language Teacher Training in Russia in the last 20 years	7
2.2. Overview of FL teacher training curriculum at university level	14
2.3. The current final assessment system for future Foreign Language teachers in Russia	23
Chapter 3. Literature Review Part I: Language Teacher Language Competence: theoretical and practical considerations	26
3.1. Language teacher language competence: general principles..	28
3.2. Language teacher language competence in national and international research projects	36
3.3. Towards a working definition of foreign language (FL) teacher language competence	43
Chapter 4. Literature Review Part II: Language Testing for Language Teachers: national and international experience	48
4.1. Language test evaluation: major dimensions	49
4.2. Direct and indirect testing	67
4.3. Language test formats	70
4.4. Language testing for language teachers: theoretical and practical considerations	89
4.5. Language examinations for English language teachers: national and international experience	94
Chapter 5. Description of the Current Final Language Examination for Future Teachers of English as a Foreign Language in Russia ...	110
5.1. Overview of documents involved in Exam design and administration	110
5.2. Content and format of the current Final Language Examination for university graduates	114
5.3. Administration of the current Final Language Exam for university graduates	119

Chapter 6.	Research Methodology	122
	6.1. Research design	123
	6.2. Empirical data collection and analysis	127
	6.3. Ethical issues of research	151
	6.4. Limitations of research	152
Chapter 7.	Findings on RQ1: design and administration of the current Final Language Examination for teachers of English as a Foreign Language in Russia	154
	7.1. Survey of the Faculty of Foreign Languages staff	155
	7.2. Post-exam survey for examiners and Exam takers	171
Chapter 8.	Discussion of findings on RQ1: procedures of Exam design, piloting and administration as seen by different stakeholders ...	179
	8.1. Final Language Examination materials design	179
	8.2. Administration of the Final Language Examination	182
Chapter 9.	Findings on RQ2: How relevant is the Examination content to the language needs of practising English teachers? What are the language needs of language teachers in Russia?	188
	9.1. Final Examination content and format	188
	9.2. Exam content as seen through the interviews of school teachers of English	191
	9.3. Needs analysis of teachers of English as a foreign language	198
	9.4. Needs analysis of the Final year university students	205
Chapter 10.	Discussion of findings on RQ2	212
	10.1. Final Language Examination content and format as seen by the stakeholders	212
	10.2. Reflection on the obtained data: how relevant is the Exam to language teacher needs?	219
Chapter 11.	Findings on RQ3: What are the strengths and weaknesses of the current Final Language Examination? What changes, if any, might be required?	223
Chapter 12.	Discussion of the Final Language Examination strengths and weaknesses as seen by different groups of stakeholders	234
	12.1. Opinions of Exam developers as a context for possible changes	234
	12.2. Suggestions for possible changes in the Final Language Examination for language teachers (university graduates)...	239
Chapter 13.	Suggestions for possible changes in the current Final Language Examination for future language teachers	245
	13.1. Possible changes in the Exam content, format and administration	245
	13.2. Implications for changes in the Final Language Examination	258
	13.3. Limitations of this study	264
Chapter 14.	Conclusion	266

List of references	271
Appendices (Volume II)	
Appendix 1.	The system of FL teacher development in Russia since 1990s 278
Appendix 2.	Sample tasks for the National Exam in English for schools in Russia 279
Appendix 3.	Sample tasks from Language Examination for college graduates (future English teachers), a British Council project in Russia 1999-2008 281
Appendix 4.	Administration guidelines for Final Language Examination for college graduates (future teachers of English as a FL) in Russia 288
Appendix 5.	Lesson evaluation framework for teaching practice at Tula State Pedagogical University 292
Appendix 6.	Final Language Examination for university graduates: a sample of reading text 295
Appendix 7.	Questionnaire for examiners and exam designers on the Final Language Examination in English for Russian university graduates (Survey 1) 296
Appendix 8A.	Post-exam questionnaire for examiners (Survey 2) 304
Appendix 8B.	Post-exam questionnaire for students (Survey 2) 306
Appendix 9.	Needs analysis of teachers of English as a FL in Tula region, Russia (Survey 3) 308
Appendix 10.	Needs analysis of final year university students of FL department (future teachers of English) (Survey 4) 312
Appendix 11.	Interview framework for teachers of English as a FL in Tula, Russia 315
Appendix 12.	Teacher interview transcripts 317
Appendix 13.	Categorization of open-ended responses in teacher interviews 342
Appendix 14.	Categorization of open-ended responses in Final Exam survey for examiners and exam designers (Survey 1) 352
Appendix 15.	Samples of alternative reading tasks for the Final Language Examination 358
Appendix 16.	Alternative task samples (listening, reading, speaking, TLA) 372

List of abbreviations used in this thesis

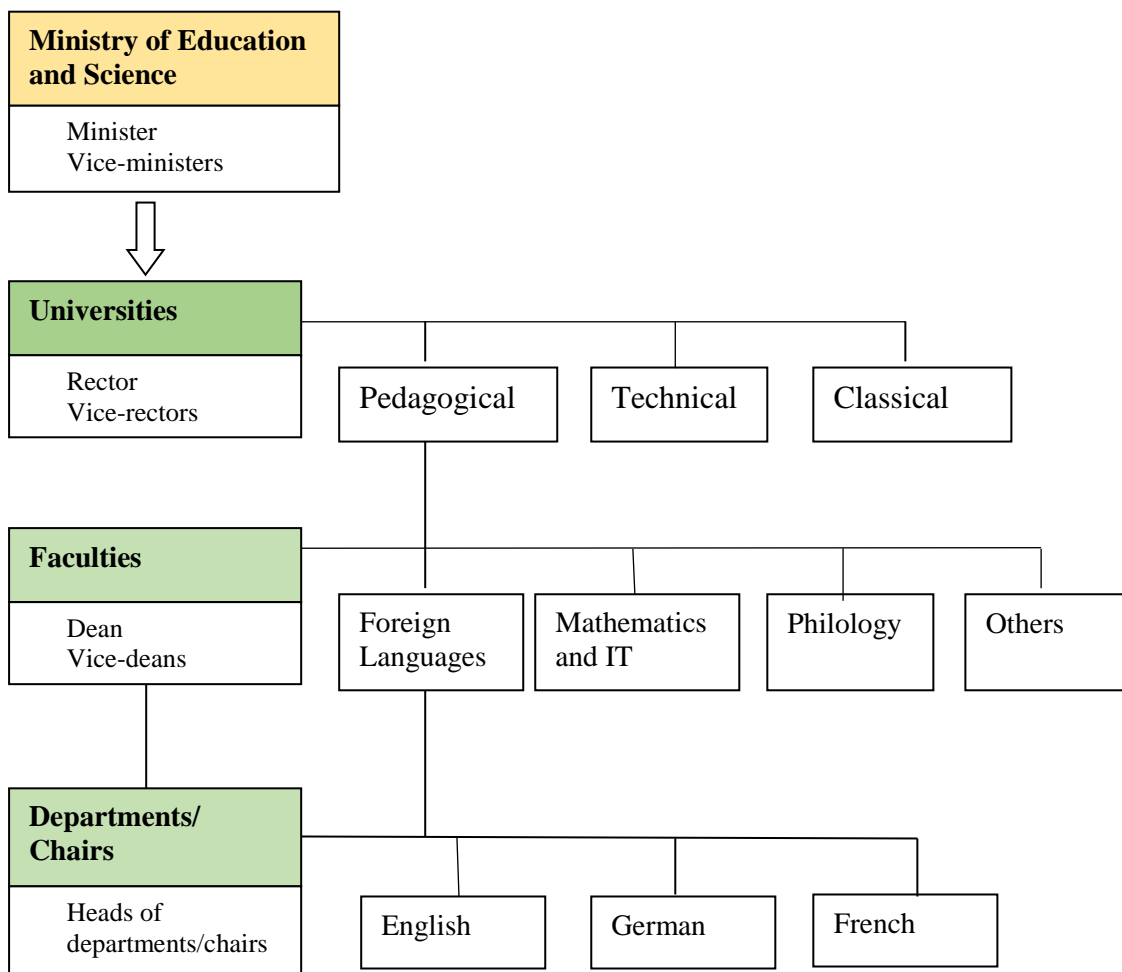
BC	British Council
Cambridge ESOL	Cambridge English (also known as ESOL) is a department of Cambridge University that deals with design and worldwide administration of international language examinations: IELTS, Proficiency (CPE), Advanced (CAE), First (FCE), Preliminary (PET) and Key (KET).
CEFR	Common European Framework of Reference (Council of Europe, 2000)
CLIL	Content and Language Integrated Learning
ELT	English Language Teaching
EPPL	Exame de Proficiência para Professores de Língua Estrangeira (Examination of Proficiency for Foreign Language Teachers, Brazil)
ESP	English for Specific Purposes
ETS	Educational Testing Service – the main standardized examination body in the USA (responsible for design and administration of internationally recognized language examinations like TOEFL and Praxis)
FL	Foreign Language
1 st FL	the foreign language (English in this research) students start learning at school and go on learning at university; the foreign language on which they take their school-leaving exam to enter university
2 nd FL	the foreign language students choose at university (starts on Year 1)
ICELT	In-service Certificate of English Language Teaching
LEA	Local Education Authorities
LPATE	Language Proficiency Assessment for Teachers of English (Hong Kong)
PRESET(T)	Pre-service teacher training
TESOL	Teaching English to Speakers of Other Languages

List of institutions, people and events referred to in the thesis

Bologna agreement= Bologna declaration	One of the main voluntary processes at European level launched in 1999, nowadays implemented in 47 states, which define the European Higher Education Area (EHEA). Members of the Bologna agreement, together with the European Commission, and the consultative members, namely the Council of Europe, UNESCO, EUA, ESU, EURASHE, ENQA, Education International and BUSINESSEUROPE. The main objective of the Bologna agreement was meant to ensure more comparable, compatible and coherent systems of higher education in Europe (http://www.ehea.info/article-details.aspx?ArticleId=5 , retrieved on June 19, 2014).
End of year/ course examination	Usually an oral form of assessment that takes place at the end of the term/course. Exam materials are usually developed by the subject teacher(s) and approved by the Chair/Department. One of exam's essential elements is an examination card. At the exam, students are always given marks from 2 (poor) to 5 (excellent). Another form of assessment is <i>credit</i> . At credits, students are exposed to shorter tasks (compared to exams) and are only assessed as pass/fail. Similarly to exams, the majority of credits are oral.

Final Language Examination	One of the two essential examinations that graduates of the FL department take at the end of their course of studies, to qualify as teachers of English as a Foreign Language
Examination card ('ticket')	A card with a number which contains questions a candidate must answer. Examination cards are printed out, approved by the Faculty Council and signed by the Dean of the faculty.
Final Examination Board	Members of the Faculty staff (usually 5-6 people) who teach different subjects – Practical course of English, Theoretical Grammar/Phonetics, Lexicology, History of Language. These are the same people who work with the students throughout their course of studies and therefore know all students well. The members are appointed by the Dean's order issued specially for the Final examination. A degree in Linguistics (usually PhD) is required.
<i>Chairperson of the Final Exam Board</i>	A professor/lecturer who is the head of the team of examiners. The chairperson is always external. i.e. works in another institution. The Chairperson is appointed by the order of the Rector's office
Final Examination Syllabus	The document developed by the Faculty which contains information about students' expected performance at the examination, topics covered, sample questions and types of texts for reading and listening.
Institutes, incl. Pedagogical	Before the 1990s, the major type of institutions of higher education in Russia with only a few universities (e.g. Moscow State University; Leningrad state University). After 1995 most institutes gained the status of universities
Ministry of Education and Science	A federal body of executive power, which elaborates state policy and normative regulation in education, scientific ...and innovative activity, intellectual property, and also in the sphere of upbringing, social support and social protection of schoolchildren and students (http://eng.mon.gov.ru/str/mon/mis/ retrieved on June 20, 2014). The Ministry of Education is the highest body in the hierarchy of educational institutions
Regional Department of Education	Highest administrative body in the region, dealing with all issues of education – from nurseries to in-service training; the Department issues laws and takes major decisions about educational programmes and school licensing. The Department is involved in licensing universities, too, although accreditation of universities is a prerogative of the Ministry of Education.
Pedagogical Universities	Universities that train only teachers; most of them are former pedagogical institutes (see above). Pedagogical universities usually have up to 15 faculties – Russian and Russian Literature; Mathematics, Biology and Chemistry; Foreign Languages; Faculty of primary education, etc.
Student/graduate/exam taker	(in this research) – a student at the end of their course of studies (5 th year – specialist or 4 th year – bachelor) who takes the Final Language Examination to qualify as a teacher of English as a foreign language.

The hierarchy of institutions in the system of Russian higher education and structure of a Russian pedagogical university



List of tables

Table 2.1:	Curriculum areas and subjects in the FL teacher training programme	23
Table 3.1:	Summary of project findings on language teacher language competence	42
Table 4.1:	Concurrent, empirical and face validity as seen by different authors	54
Table 4.2:	Major threats to language test reliability: summary	63
Table 4.3:	McNamara's view of strong and weak performance tests	69
Table 4.4:	Objective task types in language testing	74
Table 4.5:	Advantages and disadvantages of different scoring methods	87
Table 4.6:	Assessment foci and assessment tasks in language examinations for language teachers	103
Table 4.7:	International and national language examinations for language teachers: summary	104
Table 4.8:	Final Language Examination evaluation checklist	107
Table 4.9:	Design of Exam evaluation checklist and data collection	109
Table 5.1:	Representation of linguistic subjects in the topics for in Task 1 at the Final Language Exam	115
Table 5.2:	Final Language Examination: content and format	118
Table 6.1:	Data collection procedures	124
Table 6.2:	Strengths and weaknesses of empirical data collection methods ...	125
Table 6.3:	Final Exam Survey. Cross-tabulation: respondents' age*experience in the Exam	129
Table 6.4:	Final Exam Survey. Cross-tabulation: respondents' experience*role in the Exam	129
Table 6.5:	Piloting checklist for Survey 1	131
Table 6.6:	Example of categorizing qualitative responses in Survey 1	132
Table 6.7:	A sample of presentation of participant responses to Q7, Survey 1: cross-tabulation age*criteria for choosing Exam texts	133
Table 6.8:	Differences between the content of Survey 2A and 2B	134
Table 6.9:	Categorising responses to open-ended questions in Survey 2	135
Table 6.10:	A sample of data analysis for Survey 3: Cross-tabulation experience*reading ELT literature	139
Table 6.11:	A sample of data analysis for Survey 3: Cross-tabulation school type*reading ELT literature	140
Table 6.12:	Strengths and weaknesses of structured interviews	143
Table 6.13:	Structure of the interview for teachers of English	143
Table 6.14:	Teacher interview: participants' age	144
Table 6.15:	Teacher interview: participants' teaching experience	144

Table 6.16a-d:	Interview 1-4: participants	145
Tables 6.17:	Teacher interview: categorization of open-ended responses (a sample)	147
Table 6.18:	Relation of research instruments to research questions	149
Table 6.19:	Empirical data collection: summary	150
Table 7.1:	People involved in Final Language Exam materials development ...	155
Table 7.2:	Responses to Q3: <i>Who is involved in choosing Final Exam task types?</i>	156
Table 7.3:	Resources employed in Final Language Exam materials design	156
Table 7.4:	Responses to Q4: <i>Are the criteria for choosing task types clearly stated in the Exam syllabus/other document?</i>	158
Table 7.5:	Criteria for choosing task types (open-ended responses to Q4: <i>Are the criteria for choosing task types clearly stated in the Exam syllabus/other document? If yes, what are they?</i>)	159
Table 7.6:	Responses to Q5 : <i>Are there any criteria of task selection which are more important than others?</i>	160
Table 7.7:	Responses to Q6-7: <i>Are the criteria for choosing listening /reading texts presented in the Exam Syllabus/other document?</i>	160
Table 7.8:	Responses to Q8: <i>Are exam tasks moderated (scrutinized by several staff members before they become exam tasks)?</i>	162
Table 7.9:	Responses to Q9: <i>If materials/items are moderated, who is involved in the process?</i>	162
Table 7.10:	Responses to Q11: <i>Are the tasks trialled (administered to a similar group of students) before they become exam tasks?</i>	163
Table 7.11:	Responses to Q12: <i>What criteria are used by the Department in the appointment of materials designers and examiners?</i>	163
Table 7.12:	Responses to Q13: <i>Do materials writers get some training in materials design (locally or centrally)?</i> and Q15: <i>Is there any training for examiners before the examination?</i>	164
Table 7.13:	Responses to Q14 and Q16: <i>What kind of training is it?</i>	164
Table 7.14:	Analysis of open-ended responses to Q19: <i>What criteria are employed for assessing student answers?</i>	165
Table 7.15:	Responses to Q20: <i>Do all criteria have the same weight?</i>	167
Table 7.16:	Responses to Q21: <i>Are there descriptors for each criterion – what is excellent, good, etc.?</i>	167
Table 7.17:	Responses to Q23: <i>Are the existing criteria helpful in resolving disagreement (if any)?</i>	168
Table 7.18:	Responses to Q22: <i>How do examiners come to agreement about the final mark?</i>	168
Table 7.19:	Responses to Q28: <i>How long is the examination for examiners (from the very beginning to the very end)?</i>	170

Table 7.20:	Responses to Q29: <i>How much time (on average) does each student spend in the exam room (preparation time + speaking time)? Please give minimum and maximum time</i>	171
Table 7.21a:	Responses to the open-ended question ‘Examination started_____’...	172
Table 7.21b:	Responses to the open-ended question ‘Examination finished_____’	172
Table 7.22a:	Exam takers’ responses to the open-ended question ‘Time you had for preparation’	173
Table 7.22b:	Exam takers’ responses to the open-ended question ‘Time you spent answering’	173
Table 7.23:	Examiners’ responses to the question ‘Is examiner behavior specified?’	174
Table 7.24:	Examiners’ and exam takers’ responses to question ‘ <i>What were the reasons for intervention?</i> ’	175
Table 7.25:	Responses to question ‘ <i>What kind of intervention was it?</i> ’	175
Table 7.26:	Responses to question ‘ <i>What was the effect/result of intervention?</i> ’	176
Table 7.27:	Summary of examiners’ comments on the Final Language Examination	177
Table 8.1:	Examination design at university	181
Table 8.2:	Administration of the Final Language Examination at university	185
Table 9.1:	Responses to Q18: <i>In your view, what is the focus of assessment?....</i>	189
Table 9.2:	Responses to Q18: <i>In your view, what is the focus of assessment?....</i>	190
Table 9.3:	Responses to Q18: <i>In your view, what is the focus of assessment?....</i>	190
Table 9.4:	Teachers’ responses to the question ‘ <i>What activities are you involved in at least once a week?</i> ’	192
Table 9.5:	Interviewees’ opinion on the linguistic part of the Final Language Examination	193
Table 9.6:	Interviewees’ opinion on the Reading and Speaking part of the Final Language Examination	195
Table 9.7:	Interviewees’ opinion on the Listening and Speaking part of the Final Language Examination	196
Table 9.8:	Responses to Q1.1: <i>Please rate how often you are (were) involved in the following listening activities</i>	199
Table 9.9:	Responses to Q1.2: <i>Please rate how often you are (were) involved in the following reading activities</i>	200
Table 9.10:	Responses to Q2.1: <i>Please rate how often you are (were) involved in the following speaking activities</i>	201
Table 9.11:	Responses to Q2.2: <i>Please rate how often you are (were) involved in the following writing activities</i>	202
Table 9.12:	Responses to Q3: <i>Please rate how confident you feel in the following areas of the English language</i>	203
Table 9.13a:	Cross-tabulation experience*level of confidence in <i>listening</i>	203

Table 9.13b:	Cross-tabulation experience*level of confidence in speaking	204
Table 9.13c:	Cross-tabulation experience*level of confidence in grammar	204
Table 9.13d:	Cross-tabulation experience*level of confidence in ELT terminology	204
Table 9.14:	Responses to Q4: <i>As a trainee teacher at school, how often were you involved in [the following] receptive activities in English?</i>	206
Table 9.15:	Responses to Q4: <i>As a trainee teacher at school, how often were you involved in [the following] receptive activities in English?.....</i>	207
Table 9.16:	Responses to Q8: <i>Please rate how confident you feel in the following areas of the English language</i>	207
Table 9.17a:	Listening skills teachers of English in Russia require	208
Table 9.17b:	Listening skills student teachers of English in Russia require during teaching practice	209
Table 9.18a:	Reading skills teachers of English in Russia require	209
Table 9.18b:	Reading skills student teachers of English in Russia require during teaching practice	209
Table 9.19a:	Speaking skills teachers of English in Russia require	209
Table 9.19b:	Speaking skills student teachers of English in Russia require during teaching practice	210
Table 9.20a:	Writing skills teachers of English in Russia require	210
Table 9.20b:	Writing skills student teachers of English in Russia require during teaching practice	210
Table 10.1:	Involvement of professional knowledge and communicative skills in teacher activities in and out of the classroom	217
Table 11.1:	Responses to Q30: <i>What do you feel about the exam?</i>	223
Tables 11.2a-f:	Responses to Q31: <i>Please state what you think about the possible advantages of the examination listed below</i>	224
Tables 11.3a-f:	Responses to Q32: <i>Please state if the following can happen in your situation (your department); and Q33: Please state if you think the following is a problem which needs to be solved</i>	227
Table 11.3g:	Responses to Q32: <i>Please state if the following can happen in your situation (your department).....</i>	232
Table 11.4:	Responses to Q34: <i>Would you like any changes to be introduced?....</i>	232
Table 11.5:	Responses to Q35: <i>What kind of changes would you like to have?....</i>	233
Table 12.1:	Possible ways of changing the Final Language Examination	243
Table 13.1:	Assessment criteria for the current Final Language Examination ...	248
Table 13.2:	Possible effects of changes in the Final Language Examination for language teachers (university graduates)	263

List of figures and pictures

Figure 2.1:	Curriculum areas for FL teacher training at university	15
Picture 2.1:	A sample of examination card: examination on Theoretical Phonetics at the end of the course	19
Picture 2.2:	A sample of documentation (matriculation book) illustrating a student's exam marks for a semester	20
Figure 4.1:	Major types of language test validity	52
Figure 4.1:	Types of input for oral production tasks	83
Figure 4.3:	Communicative skills under assessment in ICELT	98
Figure 4.4:	Representation of communicative skills and knowledge in international language examinations for language teachers	99
Figure 4.5:	Representation of communicative skills in national language examinations for language teachers	100
Picture 5.1:	Regulatory documents for final examinations at Russian universities.....	111
Picture 5.2:	An examination card for the Final Language Examination	114
Picture 5.3a:	Examination room at the beginning of the Final Language Exam ..	120
Picture 5.5b:	Examination room from 10am till the end of the Final Language Exam	120
Figure 6.1:	Stages of this empirical research	123
Figure 6.2:	Stakeholders involved in data collection in this research	127
Figure 6.3:	SPSS database for the Final Exam Questionnaire	132
Figure 6.4:	A sample of visual presentation of qualitative responses to Q7, Survey 1.....	133
Figure 6.5:	Research population in Survey 3: levels the respondents teach at	137
Picture 6.1:	Survey 3: coding responses sent by email	138
Figure 6.6:	Statistics from Survey 3: How often teachers read ELT literature	139
Picture 6.2:	On-line survey: screenshot – multiple choice	141
Picture 6.3:	On-line survey: screenshot – scale	142
Picture 6.4:	Teacher interview: introducing quantitative data in SPSS	146
Figure 7.1:	Criteria for choosing task types (statistical summary of open-ended responses)	159
Figure 7.2a:	Criteria for choosing listening texts	161
Figure 7.2b:	Criteria for choosing reading texts	161
Figure 7.3:	Frequency of mentioning the assessment criteria by the respondents	166
Figure 7.4:	Responses to Q24: <i>How many people are there usually present in the actual examining committee?</i>	169

Figure 9.1:	Teachers' opinions on Exam Task 1: Linguistic knowledge	192
Figure 9.2:	Teachers' opinions on Exam task appropriacy (Reading/Listening and Speaking)	194
Figure 9.3:	Teachers' opinions on the choice of Exam texts	194
Figure 9.4:	Teachers' opinions on speaking skills assessed at the Exam	195
Figure 9.5:	Teachers' suggestions of Exam changes	197
Figure 9.6:	Summary of responses to the question: <i>'From your current position, what parts, do you think, the exam should consist of?'</i>	197

Chapter 1

Introduction

The research focuses on the Final Language Examination for teachers of English as a Foreign Language (FL) in Russia at the pre-service (university) level. It is a 'qualification' examination that all graduates take at the end of their programme of studies. Passing this Examination is essential for qualifying as a teacher of English as a FL. According to the available statistics, fail rates are not high, representing about 1% of Exam takers on average, with the possibility of re-sitting the Exam in next academic year. Failing the Exam means not graduating from the university and not qualifying as a teacher. The outcome of the Exam – getting a teaching qualification or failing to – place the Exam in the group of high-stakes examinations. For many university graduates, the Final Language Examination at the pre-service level is the last formal language examination they will take in their professional life. Unlike some in-service systems described in the literature (e.g. Lavigne, 2014), the current Russian system of in-service teacher development does not involve any further formal language assessment/evaluation of practising teachers, nor does it make obligatory any in-service language training. The situation may vary depending on the teaching context: in some private language schools teachers of English are expected to develop and maintain their language skills. In comprehensive schools, however, many English teachers lack good opportunities for in-service language development. Free in-service courses provided by the local education authorities, which all teachers are supposed to take every 5 years, are conducted in Russian. If such a course includes assessment, it too is administered in Russian. Thus, the role of language development of future teachers of English at university, and the role of formal summative assessment at the end of the pre-service training programme are crucial for acquiring the language skills that a language teacher needs.

This research presents a case study of Final Language Assessment in one institution of Higher Education in the Tula region. In this region of 2 million population, there are 2 universities that train Foreign Language teachers, the leading one being Tula State Pedagogical University. Students of the FL Department of Tula State Pedagogical University take two final examinations to qualify as teachers of English or another

foreign language (German, French): the Final Language Examination, which is the focus of attention in this research, and an examination on Theory of Education, Psychology and TESOL Methods¹. Both examinations are obligatory for all graduates.

According to State Educational Standards, final assessment is an entirely internal process defined and administered by a university, with no further approval required of the LEA, Ministry of Education or other authorities. Nevertheless, there is no indication in the Standards that excludes the involvement of other stakeholders, such as school teachers of English, in the process of materials design and/or piloting. Moreover, a closer look at State Educational Standards (2010) reveals that:

8.4. An institution of higher education must provide conditions for [students] assessment which are as close as possible to their future professional functioning. For this, potential employers and lecturers on other subjects [apart from the one(s) assessed] should be involved in assessment in any possible way (2010: 13).

The Final Language Examination aims to assess graduates' command of the target language (communicative skills) and linguistic knowledge that they have acquired during the course of studies. The Examination is administered orally. Exam takers are expected to answer a linguistic question and carry out one listening and one reading task. Each student (Exam taker) gets an Examination card presenting three tasks: Task 1: a linguistic question; Task 2: Reading; Task 3: Listening. In the Examination room, each student has 60 min to prepare the linguistic question from the Exam card, to read and to listen to the texts. Linguistic questions, and texts for reading and listening are different for each student. Students are allowed to take notes while preparing their answers, but the notes are not taken into consideration during the scoring process. Students answer the linguistic question and do the reading and listening tasks orally, in front of the Examination Board. Students can hear each other answering. Usually there are 10-12 students taking the Examination within one day. When all the students have finished their answers and left the Exam room, examiners discuss the final marks. Then the marks are announced to students on the day of the Examination.

This research was inspired by two issues. First, it is an attempt to bridge the gap between the high status of the Final Language Examination and its importance for teacher development, and the lack of research and changes in this area over a period

¹ The examination on the Theory of Education, Psychology and TESOL is administered in Russian

of at least 20 years. This study can be viewed as one of the initial steps in investigating the advantages and disadvantages of the current system of Final language assessment of university graduates (future teachers of English) in Russia and suggesting ways of making possible changes in the Final Language Examination.

The second trigger for conducting this research was my personal experience as a university lecturer at the Department of Foreign Languages of Tula State Pedagogical University. I have always had some questions in mind: At university, outside the language classroom, are we assessing what our students (graduates) will really need in the future? At the last stage of their pre-service teacher training, what assessment should students undergo?

As a teacher who went through a similar examination at the end of my university course and taught English for more than 10 years at different stages, I have felt that the current Examination content and format do not take into consideration the reality that graduates have to deal with. As an exam administrator at Tula Teacher Training College #1 (2000-2005) and a head of exam development team in a long term British Council project (1999-2007) I realized that there are different ways to test language development, including language testing for teachers.

Taken together, several factors informed the present study:

- importance of the Examination for university graduates qualifying as teachers of English;
- recent developments at secondary and vocational stages of education in Russia, with universities not being involved (Chapter 2, pp.10-14);
- lack of research in assessment for language teachers in Russia, and quite a considerable gap between the current Final Language Examination for language teachers and the existing national and international experience in the area.

The study is guided by the following **research questions**:

1. What are the procedures for Exam design, piloting and administration as seen by different stakeholders? This includes investigation of:
 - procedures for selecting the content and defining the format of the Examination;
 - design and choice of assessment tasks;
 - design and use of assessment criteria.

2. How relevant is the Exam content to the language needs of practising English teachers?
What are the language needs of language teachers in Russia?
3. What are the strengths and weaknesses of the current examination? What changes, if any, might be required?
4. What are the possible alternative versions to the current Final Language Examination for language teachers?

The thesis consists of the Introduction, 14 chapters, the Conclusion and the Appendices.

The *Introduction (Chapter 1)* explains the choice of research topic – Final language assessment of future English teachers (university graduates) in Russia. The Introduction starts with a brief overview of the Final Language Examination and its role in English teacher language development. Then the Introduction presents the research questions. Finally, an outline of the thesis is given together with a short description of the content of each chapter.

Chapter 2 describes the context of the research and looks into the system of foreign language teacher training in Russia and the current assessment system for trainee teachers of foreign languages at university.

Chapter 3 aims to investigate theoretical and empirical research previously conducted in different countries on language teacher language competence, its structure and constituent elements. Chapter 3 starts with an overview of theoretical considerations of language competence and language teacher competence. Then the outcomes of several projects are analyzed. Chapter 3 concludes with a suggested description of language teacher language competence which, for this research, serves as a springboard for the current Final Language Exam evaluation.

Chapter 4 looks at various parameters of test evaluation – validity, reliability, authenticity and practicality. Chapter 4 provides insight into assessment of FL teacher language competence as performed by national (Hong Kong, Brazil, USA, Australia) and international (Cambridge ESOL, ETS) examination bodies. The major outcome of Chapter 4 is a taxonomy of language skills and areas of knowledge that are the focus of assessment of different language examinations for language teachers, and a taxonomy of tasks that can be applied to language teacher competence assessment.

Chapter 5 presents a description of the content and format of the current Final Language Examination for university graduates in Russia based on analysis of federal and regional documents – State Standards for teacher development (2010), Final Examination syllabus (2010), and samples of Final Examination materials.

Chapter 6 presents the methodology of this research. It describes the key stages of the study, research instruments, and stakeholders involved in data collection. The chapter also presents the stages of research and methods of data collection and analysis.

Chapters 7-12 present empirical findings on the current Final Language Examination development and administration and also demonstrate how different stakeholders see the Examination. *Chapter 7* presents quantitative data on Research question 1, obtained through specially designed surveys for Exam developers, examiners and Exam takers. *Chapter 8* discusses those findings and summarises the key issues of Exam design, including content selection and administration.

Chapter 9 deals with the findings on Research question 2 and examines the relevance of the current Exam to the language needs of English teachers. This chapter presents the qualitative and quantitative data from teacher interviews and English teacher needs analysis performed in the Tula region. *Chapter 10* discusses the relevance of the current Exam foci and assessment tool to the activities that English teachers deal with regularly in their professional life.

Chapter 11 deals with the quantitative data from Exam designers and administrators regarding some strong and weak points of the current Final Language Exam and possible changes that the Exam might require. *Chapter 12* speculates on the obtained data, summarizes the strengths and weaknesses of the Exam, and discusses threats to the Exam's validity, reliability and authenticity. The chapter also maps out some directions for possible changes in the current Examination.

Chapter 13 looks at some practical ways of implementing changes and presents alternative Exam tasks and assessment criteria. The chapter also discusses implications of the changes for the Final Language Exam itself and assessment practices at university.

The Conclusion sums up major research findings and defines the ways this research contributes to knowledge and experience in the area of FL teacher training and assessment. The Conclusion also presents directions for further research.

Chapter 2

Context of the Research: Foreign Language Teacher Development in Russia

This chapter starts with an overview of the contemporary system of pre-service training of foreign language (FL) teachers in Russia and gives an insight into the past 20 years – the years of drastic change and development in Russian education – at secondary (school), tertiary (college) and higher (university) levels. The chapter describes teacher training programmes at college and university; the first college-university English teacher training projects in Russia (1998-2007); and the first attempt in Russia to change progress and final language assessment for future English teachers.

Part 2.1 describes innovations in the system of education in Russia which, in various ways, have influenced the process of teacher training: the National Examination for schools (<http://www.ege.edu.ru/>, retrieved 17.04.14) that has been obligatory for all school leavers since 2005; changes in the Standards of Higher education and university curricula due to Russia joining the Bologna agreement; and adoption of the ‘3 level’ model of higher education ‘Bachelor – Master – Doctor’ instead of the usual 2-level model. All these aspects help us to see the Final Language Examination as part of the teacher training system in Russia with its advantages, restrictions and possible perspectives.

Part 2.2. gives an outline of the courses that trainee FL teachers take throughout 4/5 years of study, and the continuous assessment they have to pass to become eligible to take the Final Examinations and graduate. All these are important for looking at the current Final Language Examination as an integral part of the FL teacher training programme. It is hoped that a broader overview of this kind help us to look at the strengths and weaknesses of the Final Language Examination as a part of the whole curriculum, rather than as an independent language examination, and therefore, to develop realistic suggestions for change, if required.

2.1. Changes in the Russian system of Foreign Language Teacher Training in Russia in the last 20 years

Foreign language teacher training has always been an essential part of Russian education and its importance increased when educational reforms were initiated by the Government in the 1990s – 2000s. As was stated in those years, no reforms were possible at the school level without investing in teachers, which suggested major reforms in teacher education. Later, however, in the early 2000s, teacher training became the area of least attention and interest on the part of the Ministry of Education. In contrast to lack of developments in teacher training in general and FL teacher training in particular, some changes took place at the secondary school level when the National Examination for schools was introduced and piloted in 2000, and then launched in all regions of Russia. The major purpose of the National Examination was unification of requirements for school leavers' knowledge and skills. The National Examination also aimed to provide school leavers all over Russia with equal opportunities to enter colleges and universities. First, examinations for entrance to colleges and universities were eliminated, and applicants were expected to send/present results of the National Examination to be considered by admission committees. Second, school leavers could choose several universities to apply to. This can be done by post or Internet to save applicants from travelling – a condition really important for such a big country as Russia. The National Exam for foreign languages – English, German, French and Spanish – was the first language examination in Russia which followed international requirements – from fundamentals like the concept of communicative competence with 4 language skills involved, to exam administration and marking (including multiple markers and scoring scales for oral and written parts). When first introduced in 2000 in some pilot regions, the exam caused a lot of problems – from new task types to lack of understanding of examiner behaviour. In a way, the National Examination for schools (http://www.rustest.ru/ege/ege_2013/, retrieved 23.04.14) became a milestone in educational reforms: it first of all influenced all classroom practices in secondary schools and, indirectly, changed expectations of assessment systems at the upper levels of education, namely colleges and universities. Samples of the National Exam on English are presented in Appendix 2.

The task of FL teacher development was traditionally fulfilled by pedagogical institutes and universities. Since the 1960s, pedagogical institutes, many of which gained the status of universities in the 1990s, trained teachers with courses of studies lasting 5 years. In 1995, Tula Pedagogical College №1 started its own FL teacher training programme. Lacking sufficient teaching staff, the college, with the support of Local Educational Authorities (LEA), recruited specialists from Tula State Pedagogical University. 1995 was the first year in which school leavers in Tula who wanted to become teachers of English were able to choose between entering the pedagogical college or the pedagogical university. In 1998 the British Council started the 'Fast-track English teacher training' project in Russia, in the Tula Pedagogical College №1, which by that time had developed an experimental programme (syllabus and teaching materials, including the first English coursebook for colleges in Russia (Malchenko, Okninskaya, Lyubimova, 1996) but still remained the only college in Russia which trained teachers of English. The aim of the BC project was to develop the system of English teacher training at pedagogical colleges which would, in addition to the efforts of universities, help to cover the shortage of English teachers at schools², being more practice-oriented and faster than the university model, since instead of 5 university years the college course was designed to last 3 years.

In 1999, after the first cohort of students graduated and 2 more cohorts were admitted to the College, the project proved to be successful: all college graduates passed their final examinations, and all were employed as English teachers in Tula and the Tula region, receiving good feedback from senior colleagues and school administrations. It was a big achievement by the college team because, apart from a good quality English teacher training programme, the college managed to change society's view of colleges as non-prestigious educational institutions. Soon, by 2003, more colleges all over Russia started FL teacher training programmes. In 2005 there was a network of 100 colleges which trained FL teachers: mostly teachers of English as an FL but also, in some regions, involving other languages – Polish and Lithuanian in the Western part of Russia (Kaliningrad region), and Finnish in St.Petersburg; while some colleges in the Far East planned to introduce Chinese as a second FL. Thus, from 1996 in Tula and

² The shortage in the 1990s was caused by a large number of FL teachers leaving the job for other, better paid ones (as interpreters, translators, secretaries in international companies, etc.)

2000 in other parts of Russia FL teacher training was conducted by both pedagogical colleges and universities; the applicants could choose between a 5-year programme at university (higher professional education) and a 3-year programme at college (tertiary professional education), as presented in Appendix 1.

One of the project's objectives was development of a new format of English examination for future teachers of English which included progress and final examinations for college students. The work included a number of college staff training in Russia and in the UK (Lancaster University, IELE; University of Warwick; University of Plymouth, University of Leeds). The first step was to change the examination format³ bringing it closer to the Common European Framework (CEF) requirements and the National Examination for schools in Russia, and to international language examinations (e.g. Cambridge ESOL examinations) in terms of skills assessed, tasks employed, and papers contained in the exam. The content of the college examination was revisited, too, in an attempt to embrace both general and teacher-specific topics and text-types⁴. The administration procedures part was completely revised with many changes introduced – from distribution of roles at the examination to a new system of assessment criteria which required substantial examiner training⁵.

The majority of college graduates preferred to go to university to obtain higher professional (pedagogical) education, which gave them the right to work with various age groups and at different stages of education (from primary school to the teaching of advanced English courses to adults). In 2000 it became obvious to all the parties involved – the college team, local education authorities, and the university staff, that universities and colleges should work together to provide college graduates with such an opportunity. As a result, a new college-university model of English teacher training was first developed in Tula and then disseminated to other regions (Krasnoyarsk, Omsk, Volgograd, Yekaterinburg), and was projected as part of a new British Council project in Russia (Malchenko, 2005; Sokolova, 2005). It presupposed three years of college training and three years at university, the students spending most of their time

³ The changes resulted in the whole system of marking and grading being revisited, as the official written part was introduced for the first time in many years in addition to the traditional oral form which was also revised and reshaped

⁴ See Appendix 3: Sample tasks from college English Exam for future language teachers

⁵ See Appendix 4: Administration guidelines for college English Exam for future language teachers

at school with only a limited amount of contact hours and examinations at the Foreign Languages Department of the university. It eventually became obvious that the new assessment system at colleges collided with the old assessment system at universities which preferred to stick to ‘the old’ format of oral examination with examination cards (экзаменационные билеты) developed years earlier⁶. The current Final Language Examination under study has been based on a similar format. Thus, development of a new assessment format for universities, bringing it closer to the one at colleges on the one hand and the National Examination for schools on the other, was essential. It would provide coherence in assessment methods at all 3 levels of education: school, college and university.

In 2007 the Ministry of Education and Science of the Russian Federation abandoned the English language teacher training programme at colleges and the majority of them were closed. No official explanation was given but one of the reasons might have been population decrease – there were not enough school leavers for both colleges and universities. The ‘college-university’ programme was abandoned also, together with development of the new assessment system. Thus, since 2008, FL teacher training has been carried out only at pedagogical universities or pedagogical departments of ‘classical’ universities. This indirectly allowed universities to keep the same exam formats they had followed for years.

The Final Examination format has always been defined by universities. This right is given to them by the State Educational Standards which all universities must follow. Before 2010, the State Standards for FL teacher training, as well as other standards for universities, were revised by the Ministry of Education and Science on a routine basis, i.e. every five years. In 2010, a new ‘generation’ of Standards was published, representing the first attempt of the Ministry of Education to adjust Russian educational standards to Bologna requirements. Russia joined the ‘Bologna club’ in 2003, when the agreement was signed. The agreement presupposed integration of the Russian educational system and the systems of 39 other countries - members of the Bologna group (Neave, 2003; Huisman, van der Wende, 2004). The major purpose of

⁶ The data were obtained within the British Council project from universities in different parts of Russia (Central Russia, North-West region, Siberia, the Volga region). Differences in the examination format and content were observed but they were minor and did not influence the main ‘exam card’ principle

joining the Bologna agreement was, as stated by the Russian Ministry of Education, to give Russian university graduates the opportunity to continue their studies in European universities (Гретченко, 2006). Quite a number of Russian universities remained skeptical about such perspectives, but this skepticism resulted in a very limited number of publications (Акулич, 2005; Гретченко, 2009, 2012; Иркутская, 2011). The major points of criticism were:

1. graduates from previous years (before 2003) were admitted to many universities all over the world without having internationally recognized diplomas;
2. the percentage of those who go abroad for further studies was still too small to change the whole system of higher education;
3. apart from having appropriate degrees to enter a European university, Russian applicants still have other requirements to meet: international tuition fees, visa status, etc.

Another motive for joining the Bologna agreement, discussed mostly in relation to key Russian universities like Moscow State University and St. Petersburg State University, was to make Russian higher education more attractive for applicants from other countries (Order of the President of RF №599 from 7.05.12: <http://минобрнауки.рф/проекты/ведущие-вузы/мировые-рейтинги>, retrieved 25.04.14) by making Russian diplomas of higher education internationally-recognised. This was supposed to contribute to an increase in international students at Russian universities, in student exchanges, and in visiting lecturers and, in the long run, to active co-operation between Russian and foreign universities.

For Russia, joining the ‘Bologna process’ meant, first of all, big changes in the whole system of higher education. The biggest and most difficult step was switching from 5-year programmes which trained ‘specialists’ – engineers, teachers, economists, etc. - to 4-year baccalaureate programmes with 2 more years for Master’s degrees for some graduates. As often occurs during transition periods, many universities took the opportunity to stick to the ‘old’ 5-year programmes of teacher training. In 2005-2010 this resulted, and sometimes still results, in at least 2 cohorts of students included in the same department (Irkutskaya, 2011): those who entered a 5-year programme to gain the teaching qualification and those who entered a 4-year Bachelor’s programme (Appendix 1).

Apart from changes in Educational Standards and changes in the content of educational programmes at universities, adopting the two-level system meant:

- introduction of the European Credit Transfer System, that is changing the assessment system;
- introducing the system of university ranking according to established criteria (<http://ranking.ntf.ru/>, <http://ranking.ntf.ru/p139aa1.html>, retrieved on April 16, 2014).

In accordance with the Bologna regulations, the structure of the Standards was changed (Federal Law of 1 Dec 2007 #309-Ф3 ‘On introducing amendments into separate legislative acts of the Russian Federation’ (in the section concerning the change of concept and structure of the State Educational Standards) together with the approach to defining the content of educational programmes (www.ranking.ntf.ru, retrieved on April 21, 2014).

Besides freedom to define the aims and content of the training, each university gained the freedom to choose the format and content of Final examinations, including language examinations. This resulted in universities sticking to old formats which had been employed for up to twenty years and in some cases in excluding from the curriculum the Final language examinations which had formerly been part of the ‘obtaining teacher qualification’ scheme⁷. Universities’ freedom to define assessment frameworks can be seen as a possible reason for not changing old assessment systems even if they needed to change. However, such freedom can also be a good opportunity for those universities that want to develop and pilot new assessment models which would work in their own context.

The system of higher professional education in Russia is gradually approaching the point of involving various stakeholders in defining the aims of educational programmes and taking part in evaluation of their outcomes. Nowadays, universities have quite a number of supervising bodies:

- National Fund for Specialist Development (НФПК, www.ntf.ru),
- Rosobrnadzor (Accreditation institution www.obrnadzor.gov.ru),
- Federal Institute of Education Development (ФИРО, www.firo.ru).

⁷ A more detailed description of the situation throughout Russia lies outside the research aims and area

However, their role is limited to observing that the requirements of the State Educational Standards issued by the Ministry of Education of Russia are met by each university. Their function does not entail influencing or defining the format and/or content of either progress or final assessment: these are the responsibility and prerogative of the universities. There is, however, another group of stakeholders – potential employers – who are now encouraged to be involved in student progress assessment throughout the course of studies and Final Assessment⁸ (State Educational Standards, 2010: 13).

In 2008-2012 there was an attempt to introduce external evaluation in Russian Higher education. The National Accreditation Agency (www.nica.ru) designed and administered internet tests in many subjects for university students (Federal i-test in professional education). These tests became obligatory for university accreditation. The tests in English and other FL languages were piloted once and were criticized by FL departments because they:

- included only multiple choice items with some distractors not fulfilling their function;
- tested only reading and writing (spelling);
- were knowledge-, and not competence-based.

The criticism resulted in a number of publications (Soldatkin, 2003; Kuzmina, 2004; Kuzmina, Sternina, 2009) which could have lead to improvements in the area. But instead, the external evaluation project was closed in 2012 by the Ministry because ‘from 1 Jan 2012 external assessment was no longer a part of university accreditation’ (<http://www.nica.ru/accred/algorithm/>, retrieved 16.04.14) due to ‘the tendency to move the responsibility from external evaluation bodies to internal self-evaluation done by universities’ (www ranking.ntf.ru retrieved 21.04.14) as a follow-up to joining the Bologna agreement.

Taking into consideration the development of the FL teacher training system in Russia it can be stated that:

- English (and other foreign) language teacher training in Russia takes place at the university level only;

⁸ This requirement is not observed in the situation under study

- only fragmentary research has been traced in the area of *FL teacher language development* (Sokolova, 1999; Kryuchkov, 2003; Shchukina, 2008; Gubzhokova, 2010): involving communicative skills development and sociocultural development of future teachers of English;
- no evidence in the form of Ministry documents, institutional and individual research publications, or published reviews of literature has been traced in Russia which would deal with language aspects of FL teacher assessment and evaluation (including the Final examination). This statement is based on the results of a search of printed and online publications on English language teaching and the teacher development area: the ‘Foreign Language at School’ journal (2004-2014); publications of Moscow State Pedagogical University and St.Petersburg State Pedagogical University; publications of Moscow State University (FL department) and Voronezh State University (department of Roman and German Philology).

2.2. Overview of FL teacher training curriculum at university level

Teacher training programmes, including those for training FL teachers, are designed by universities. The document they must follow is the State Educational Standards issued by the Ministry of Education every 5 years. Before 2010, Standards were very precise and specific: they prescribed the minimum number of academic hours⁹ for each discipline (subject) and the content of each discipline. The version issued in 2010 and currently employed, very often referred to as ‘Generation Three’ Standards, differs a lot from its predecessors: it still contains recommendations for distribution of academic hours between courses but there is no detailed description of either content or skills that students are expected to develop. All these became the universities’ prerogative.

⁹ ‘Academic hour’ is the main measurement unit which helps to define the length and intensity of a course/discipline. Academic hours include contact hours in the classroom and students’ independent work. For example, History of English may be 72 hours long, i.e. 2 hours a week on average, whereas the Practical Course of English may take 600-900 hours.

When developing their curricula, universities are now expected to take decisions on every aspect of teacher training – from overall approach to teaching to the number of hours allocated for each subject, the content of each subject, continuous assessment, final assessment, number of teaching practices, etc. Whatever decisions are taken by programme developers, according to the Standards, the curriculum for FL teacher training must include 3 areas (Figure 2.1). Areas shown in grey are the same for all departments of the university; that is future teachers of English, future teachers of Biology/Mathematics, etc. are exposed to the same content, same assessment and approximately or exactly the same amount of hours. Specific areas are shown in white.

Figure 2.1. Curriculum areas¹⁰ for FL teacher training at university

Curriculum areas			
General <i>Humanitarian and socio-economic subjects (ОГСЭ)</i>	Professional <i>General professional development (ОПД)</i>	Linguistic <i>Professional development (ДПП)</i>	Teaching practice
History;	Psychology (Age psychology; Educational psychology; General psychology);	English (semester 1-8/10);	Lesson observation (semester 5); Conducting English lessons (semester 8, 9)
Russian;		English grammar (semester 1-6);	
Sociology;	Pedagogy (Theory of Education);	English phonetics (semester 1-6);	
Basics of Economics;		Theoretical grammar (semester 8);	
Anatomy;	Theory of education for special needs children;	Theoretical phonetics (semester 5);	
First Aid;	TESOL methods	Lexicology (semester 6);	
Basic Medicine			History of English (semester 4); Stylistics (semester 8)

Only one of the four areas – the **Linguistic area** – is taught in English. TESOL courses have traditionally been taught in Russian, although some exceptions are possible, for example, in Tula in 2005-2010. The academic year 2003-2004 was the first in which the TESOL course was taught in English at the FL department of Tula State Pedagogical University. The reasons for this were, first of all, changes in the methods of teaching English in Russia: the 1990s-2000s were the time of transition from

¹⁰ Translation of the Russian terminology is used here. Russian terms are presented in parentheses

grammar-translation to communicative methods. For various reasons the majority of publications in the 1990s (Cambridge ELT series, Longman, Macmillan, OUP) and events (e.g. those sponsored or organised by the British Council or the American Center) were available only in English. This caused quite considerable difficulties for language teachers who graduated in the 1970s-80s and who had their TESOL courses taught in Russian, because those generations of teachers had been exposed neither to ELT terminology, nor to reading or discussion of ELT issues in English. These were not just language problems, but were mostly conceptual difficulties that arose when an English term meant nothing for readers/listeners. For example, many teachers could not understand what ‘reading for gist’ meant, as reading had always been treated only as ‘reading for detailed understanding’¹¹. A new generation of English coursebooks that appeared as a result of changes in methods of English teaching (New Millennium English, 2002, <http://www.newmillenniumenglish.ru/index.php?lang=en>; Enjoy English, 2000; Millie, 2005) and were mostly a joint effort of Russian and British coursebook writers, included Teacher’s Books, Resource Packs and links to web-pages in English. All these factors contributed to changes in the ‘TESOL methods’ syllabus, and the ‘TESOL methods’ course at Tula State Pedagogical University was launched in English in 2005-2010 as an experiment which I co-ordinated with another member of staff involved in it. Lectures and seminars were delivered in English, suggested reading lists contained both English and Russian sources, and continuous assessment was done in English, although students and lecturers could always use Russian if/when they felt it was necessary, especially when discussing the peculiarities of teaching English to speakers of Russian. Special emphasis was laid on development of ‘teacher technical vocabulary’ in English and Russian, because it was obvious that graduates would require both English and Russian terminology to carry out their tasks at school.

In addition to changing the language of instruction, the assessment system for the subject was first reconsidered in 2007, with major changes taking place in 2008-2010. Instead of exam cards with questions, the exam consisted of a computer test on TESOL concepts, plus lesson planning in which students demonstrated their practical skills. Design of close-ended test tasks instead of series of exam questions was in some way

¹¹ Information provided mostly by INSET lecturers who worked with Russian teachers of English in the 1990s

influenced by the appearance of the Cambridge ESOL Teacher Knowledge Test® (TKT) which demonstrated one of the ways of assessing English teacher knowledge. The computer format was chosen because it was at this time that the university implemented 'Moodle', a virtual learning environment platform that provided a good opportunity to design computer tests for different subjects. Such substantial changes in the TESOL exam format represented an attempt to find an alternative to 'traditional' exam questions administered in Russian, which mostly presupposed reproduction of TESOL coursebook chapters. Other alternatives, like problem-solving tasks and evaluation of teaching materials were also considered. The new exam format for the TESOL, along with its content delivered in English, were supposed to enlarge the amount of language practice for trainee teachers and develop some teacher language skills that the Practical Language Course lacked.

At the end of the TESOL course, around 80 students were offered an anonymous questionnaire that aimed to evaluate the course's efficiency, difficulty and usefulness for their future jobs as English teachers¹². Students were also asked whether they found the course more difficult because it was taught in English, and whether they thought the course should be taken in English or in Russian. The greater part of the respondents marked the course efficiency as 'high'. Almost 90% of students, somewhat unexpectedly, stated that the course should be taught in English even though it was more difficult, especially at the beginning. Almost 90% of the respondents were in favour of a computer test, as opposed to an oral examination with examination cards. The major reasons for the choice of a computer test were that such a test is more objective and marking is transparent. This support for a computer test was rather unusual because Russian academic discourse has traditionally been oral, and there is a strong belief that assessment can only be effective by means of face-to-face interaction between exam taker and examiner. Although the outcomes of the TESOL course were quite satisfactory, according to exam results and students' feedback, since 2010 the TESOL course has been taught in Russian because many teaching staff members considered it more convenient and more efficient in the Russian educational context. Thus, the opportunity for the TESOL course tasks to supplement tasks in the general

¹² By the time the course ended, students had already accomplished two rounds of teaching practice

language courses (including assessment) and contribute to the development of both general and professional English was not put into practice.

As can be seen from Figure 2.1, some subjects are taught within an academic year, or even a semester, while some, like the Course of English, are studied from the first to the last year of training, although the number of academic hours can differ from year to year.

According to the available syllabi for Theoretical Phonetics, Theoretical Grammar, Lexicology, and History of English, designed by Tula State Pedagogical University, the linguistic subjects mostly aim at developing quite extensive linguistic knowledge and skills in analysing linguistic units¹³ that can also be traced through assessment tasks. ELT-related issues are presented scantily or not at all. The Course of English, sometimes called ‘The Practical Course of English’ as opposed to the theoretical subjects referred to previously, is the subject taught throughout all years of study. It aims at developing communicative skills in Listening, Reading, Speaking and Writing, building vocabulary, and practising grammar and pronunciation skills of future teachers of English.

Each subject (e.g. Practical Course of English; Theoretical Grammar, Lexicology, etc.) presupposes students taking an exam at the end of the course and also doing a series of continuous assessment tasks throughout the course. Such tasks may include:

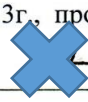
- a written test;
- answering a series of open-ended questions or problem solving (in writing or orally);
- project work (individual or in a small group).

Examinations on each subject depend on their content and aims. For example, for theoretical subjects (e.g. Theoretical Phonetics, Lexicology, etc.) it usually involves answering questions on linguistic theory and problem solving (see Picture 2.1); for the Practical Course of English, exam tasks tend to be more communicative and practice-oriented, such as listening to or reading a text and enlarging on the issues it covers, speaking on a suggested topic or describing a picture. Examinations are always oral,

¹³ Based on analysis of syllabi developed by the FL department of Tula State Pedagogical University

but students are allowed to take notes while preparing their answers. Examination tasks are presented in the form of exam cards (экзаменационные билеты). Students usually have 30-60 minutes for planning and preparing their answers in the exam room and then answer all questions orally to examiner(s). At the examinations, there is usually one examiner: the teacher/lecturer who taught the subject within a semester/academic year. Sometimes a course supervisor can take part in the examination, but that person's role is usually limited to observing whether the exams follow the guidelines designed by the Department. The examiner can ask students additional questions or give clues in case of difficulty, but these are not obligatory. The examiner is the person who listens to students give their answers and the one who does the marking. Picture 2.1 presents a sample examination card for the examination on Theoretical Phonetics¹⁴. There are 2 topics in the card, marked as 1 and 2. They aim to assess knowledge, whereas Rubric 3 is considered a practical task.

Picture 2.1. A sample examination card: examination on Theoretical Phonetics at the end of the course

<p>Министерство образования и науки Российской Федерации ФГБОУ ВПО «Тульский государственный педагогический университет им. Л.Н.Толстого»</p>	
<p>Курсовой экзамен</p>	
<p>Факультет иностранных языков</p>	<p>Дисциплина Theoretical Phonetics 3 year</p>
<p>Билет № 1</p>	
<p>1. Phonetics as a linguistic science. Theoretical and applied Phonetics.</p>	
<p>2. Word Stress in English, its nature and functions.</p>	
<p>3. Define the syllabic structure of the following words: <i>reader, kitten, pound, just</i></p>	
<p>Билет рассмотрен на заседании кафедры «<i>Я</i>» <u><i>12.</i></u> 2013г., протокол № <u><i>4</i></u></p>	
<p>Зав. кафедрой </p>	

¹⁴ Questions/tasks on Theoretical Phonetics are on the list of Linguistic questions (Task 1) at the Final Language Examination (Chapter 5, p.117)

Marks (from 2 ‘unsatisfactory’ to 5 ‘excellent’) are given by the examiner as soon as the student finishes his/her oral answer. Marks are put in students’ matriculation books (зачетная книжка) and signed by the examiner(s), as shown in Picture 2.2. Each academic year contained 2 examination diets: in winter (semesters 1, 3, 5, 7, 9) and in summer (semesters 2, 4, 6, 8, 10). During each examination diet (экзаменационная сессия) students take 2-5 examinations, only some of them being examinations on Language/Linguistic subjects and others belonging to the General and Professional curriculum areas.

Apart from examinations, assessment for some subjects from the Professional and Linguistic areas includes written coursework (курсовая работа). Students are expected to submit one piece of coursework on Theory of Education and/or Psychology (Year 3, 5th semester) and TESOL methods (Year 3, 6th semester) and one on Linguistic area subjects (Year 4, 7th or 8th semester). Coursework is essential for all students and entails a written paper (usually within 10000 words) and its oral presentation. Both the paper and the presentation are done in Russian. TESOL coursework may include samples of tasks, lesson plans, and visual aids done in English. Linguistic coursework usually presents examples of language use from fiction, periodicals or other sources.

Picture 2.2. A sample of documentation (matriculation book) illustrating a student’s exam marks for a semester

2-й семестр 20 ____ / 20 ____ учебного года

Результаты промежуточной аттестации (экзамены)

№ п/п	Наименование дисциплины (модуля), раздела	Общее кол-во час./з. ед.	Оценка	Дата сдачи экзамена	Подпись преподавателя	Фамилия преподавателя
1.	История	144/4	удовл.	7.07.14	X	X
2.	Практикум по грамматике		удовл.	03.10.14	X	X

Студент (курсант) _____ переведен на _____ курс.

The picture shows a list of exams a student passed in the 2nd semester: 1) History and 2) Practical Grammar

marks

date of exam

The fourth curriculum area which stands a bit apart from the three previously referred to is **Teaching Practice** (Year 3-4/5), when students first observe and then conduct English classes at primary and secondary school. Unlike examinations on the subjects described earlier, assessment of Teaching Practice is mostly continuous, with students maintaining specially designed diaries. A Teaching Practice diary, in some way, resembles a workbook including tasks on Theory of Education, Psychology and TESOL methods. An attempt was made in 2000 - 2001 to design a diary especially for the Department of Foreign Languages, in which TESOL tasks would be presented in English (Malchenko, Sokolova, Romashina, 2000; 2001). The major aim of such a diary was to contribute to students' grasp of ELT terminology and Classroom English and to encourage students to write (mostly lesson planning) and talk (pre- and post-lesson conferences) about classroom-related issues in English. The Diary was seen as a logical development of the TESOL course that in 2003 - 2010 was conducted in English. Nevertheless, the idea was not supported by the University. The major reason was that the diary's TESOL section completed in English, would not be 'accessible' to colleagues from the Education and Psychology departments who would then not be able to assess students' progress.

Apart from keeping a diary during Teaching Practice, students are expected to plan and conduct English classes and extra-curricular activities. The classes are observed by mentors at school and representatives from the FL Department of the university. Each lesson is discussed and a mark is given. Mentors fill in an evaluation form which serves a dual purpose: giving an outline of lesson evaluation criteria and providing a framework for post-lesson discussion and reflection. The framework is presented in Appendix 5. The major focus areas for evaluation are:

- clarity and transparency of lesson aims and whether the aims are achieved in the lesson;
- lesson structure;
- appropriateness of tasks and techniques employed;
- ways of presenting and practising language items;
- error correction;
- managing the classroom and maintaining discipline;

- the language employed by the teacher (classroom language, explaining language items, giving examples, etc.)

Post-lesson conferences are usually conducted in Russian, although there are no documentary restrictions on the choice of language. Teaching issues at post-lesson conferences can be discussed in English, if conference participants are all teachers of English. However, if representatives of the university's Theory of Education or Psychology department take part in a post-lesson conference, the discussion is conducted only in Russian. So, teaching practice does provide trainees with an opportunity to use English in the classroom and for lesson planning but pre-and post-lesson discussion is mostly conducted in Russian.

Table 2.1 summarizes key information about all three curriculum areas and teaching practice: content (for linguistic disciplines only), presence/absence of teacher-related issues, and language of instruction. This is important for further description and evaluation of the Final Language Examination which, as a final stage of the course of studies, is a reflection and logical development of the whole teacher training programme at university.

As can be seen from Table 2.1 (p.24), Russian prevails as a language of instruction in most curriculum areas – from the General area to Teaching practice. Therefore, the Final Examination in these subjects is administered in Russian. This means that the rubrics/questions are presented in Russian, students do the tasks in Russian, and additional questions from the members of the Examination Board are presented in Russian; i.e. this examination provides no evidence concerning exam takers' level of FL competence.

English is a language of instruction for theoretical linguistic subjects – Theoretical Phonetics, Theoretical Grammar, Lexicology, History of English. However, the coursework (курсовая работа) on those subjects, which is usually within 10 000 words, is done in Russian.

The Final Language Examination under consideration embraces 2 areas: linguistic subjects (Task 1) and the Practical Course of English (Task 2, 3). The content of the TESOL course, Psychology and Theory of Education do not inform the content and format of the Final Language Assessment.

Table 2.1. Curriculum areas and subjects in the FL teacher training programme

Curriculum area	Subject	Language of instruction and assessment	Professional dimension in the content	Notes
General	History; Russian; Sociology; Basics of economics;	Russian	No; aim mostly at developing general awareness (economics, sociology) or enlarges knowledge students get at secondary school (World History; History of Russia)	Out of research focus – not a subject area for assessment at the Final Language Exam
Professional	Theory of Education; Psychology	Russian	Yes	Not a part of the Final Language Exam
	TESOL	Russian/English	Yes	
Linguistic	English; Lexicology; Theoretical grammar; Theoretical Phonetics; Stylistics	English	Partially, for the Course of English (through the choice of topics) No for Linguistic subjects (see Figure 2.1)	The major assessment area at the Final Language Exam
<i>Coursework</i>	TESOL;	Russian for the body of the paper; English for examples or task samples	Yes	
	Lexicology/ Theoretical grammar/ Theoretical Phonetics/ Stylistics		No	
Teaching practice		Russian for pre-and post-lesson conferences; English for conducting classes	Yes	Not a part of the Final Exam

2.3. The current final assessment system for future Foreign Language teachers in Russia

Final Examinations are the last stage of the teacher training programme at university. Final examinations are obligatory for all students. To be eligible to take Final exams, students must pass all examinations, tests and other forms of continuous assessment prescribed by the curriculum (Year 1-4/5) with at least satisfactory marks, i.e.

minimally accepted performance¹⁵. The marks that students get at the Final examinations do not depend on marks obtained throughout the course of studies; i.e. a student who was getting good and satisfactory marks during the course of studies can get an excellent mark in the Final exam; and, conversely, a student with many excellent marks can get a good or lower mark in the Finals.

According to the current version of State Educational Standards, the number of final examinations that each graduate must take is defined by the universities. Universities also decide what examinations must be taken. Before 2010, graduates of the Foreign Languages Department took 3 Final examinations: 1st Foreign Language (English); 2nd Foreign Language (German or French); Theory of Education, Psychology and TESOL. Graduates were also supposed to present a dissertation (Выпускная Квалификационная Работа) on either Theory of Education, Psychology and TESOL (approximately 10% of graduates) or Linguistics (up to 90% of graduates). After 2010, having been given more freedom in the design of the final examination, some universities preferred to stick to earlier decisions, so the number of final examinations did not change. In Tula State Pedagogical University graduates take Final exams in:

- both foreign languages (English and German/French)
- Theory of Education, Psychology and TESOL methods;

Apart from taking these examinations, the graduate presents a Dissertation. Almost 100% of graduates write dissertations on linguistic issues. Examinations on the 1st and 2nd foreign languages, plus the examination on Theory of Education, Psychology and TESOL are regarded as 3 different exams, with a separate mark from 2 (poor) to 5 (excellent) given for each. Each exam is administered by a separate examination board. Marks for one examination do not influence marks for another. All 3 examinations have equal weight and all 3 marks are presented in academic transcripts (приложение к диплому).

The Final Examination on the 2nd foreign language (German or French) is very similar in format to the Final Examination in the 1st foreign language, so these two Final Examinations could have been investigated together in this research. However, due to

¹⁵ The marks students may get are on the following scale: 2 (poor/unsatisfactory) - 3 (satisfactory/minimum pass level) - 4 (good) - 5 (excellent)

some variation in content and format on the one hand, and limitations of this thesis on the other, the Final Examination on the 2nd Foreign Language is viewed as a separate examination and is considered a possible step in further research.

The Dissertation, whether on a TESOL or Linguistic issue, is written in Russian; i.e. the whole paper from cover and introduction to the literature list is submitted in Russian. The examiners, who are always internal, provide their feedback in Russian, the oral presentation is given in Russian, and candidates answer examiners' questions in Russian, as well. A dissertation always contains sample tasks or lesson plans for TESOL papers or samples of oral and written language use for linguistic papers, which are presented in English either in the body or in appendices. Because of these elements, the dissertation is unlikely to be considered as a means of assessment of students' written and oral performance in the 1st foreign language. Thus, the focus of this research is entirely on the Final Language Examination (English) for university graduates.

Literature Review Part I: Language Teacher Language Competence: theoretical and practical considerations

Training teachers of modern foreign languages has always considered language development of future teachers as one of the aims of instruction, especially if trainee teachers are not native speakers of the language they are going to teach (e.g. Medgyes, 1999; Coniam, 2002; Sešek, 2007; Coniam, 2013). Nowadays, aims of training are often presented in terms of competences that a learner is expected to develop (e.g. Waystage level, 1998; Threshold level, 1998; Common European Framework of reference for modern foreign languages, 2001). Professional competence of teachers in general and FL teachers in particular is an issue where a consensus has not yet been reached. The task of describing and structuring professional competence is seen as complex by many researchers due to a variety of knowledge areas, skills and attitudes that a professional is expected to demonstrate (Didi, Fay&Klaft, 1993; Anderson& Marshall, 1994; Barryman&Bailey, 1995; Tomlinson&Saunders, 1995; Nijof, 1998). Language competence of a foreign language teacher, being a part of professional competence, is a term that is quite widely used in documents (e.g. Program Standards for the Preparation of FL Teachers (USA, 2002); Standards for Teachers of Indonesian (Australia, 2005); State Standards for Teacher Development (Russia, 2010)).

Much has been written about FL teacher training and development (e.g. Ur, 1991; Parrott, 1993; Nunan&Carter, 2001; Harmer, 2001; Harmer, 2007; Harmer, 2008; Scrivener, 2005; Richards, 2002; Richards, 2010), and 'teacher competence' has become a widely used term in documents mentioned above (e.g. standards and national curricula for language teacher training in various countries). However, there is a considerable gap in TESOL and TEFL literature as far as the description of teacher language competence is concerned. As Medgyes pointed out in 1999, professional literature 'teems with books on the language learner, but is very slim on the language teacher' (1999: 21). Still, there have been a number of publications in this field (e.g. Kennedy, 1983; Thomas, 1987; Wright&Bolitho, 1993; Cullen, 1994; Cullen, 2002; Trappes-Lomax, 2002; Sešek, 2007; Richards, 2010). There may, nevertheless, be some reason to agree with Trappes-Lomax (2002) who wrote:

'There is a gap between books about language (for students, teachers, linguists) which do not deal specifically with teacher education and books about teacher education which do not deal with language' (2002: 1)

The publications mentioned above, from standards and other documents to books and articles, raise the issue of the language that a language teacher needs for effective functioning. All reviewed authors, directly or indirectly, emphasize that the command of English for an English language teacher is different from the English for people of other occupations in terms of amount of knowledge about (the) language, range of language skills, degree of accuracy and fluency of oral and written performance. Yet, as demonstrated further in this chapter, none of the authors tends to be specific describing language teachers' expected/desirable knowledge and skills. The only exception is a definition of a FL teacher linguistic competence, or Teacher Language Awareness (TLA), as it is often called (Wright&Bolitho, 1993; Widdowson, 2002). Description of 'teacher language' is considered crucial for this research because it provides a basis for the Final Language Exam evaluation, i.e. its content, format and assessment procedures and their relevance to what teachers are expected to do in the language classroom.

The literature review is divided into two parts. The current part, Chapter 3, focuses on language teacher competence, aiming to describe knowledge, skills and attitudes that a modern teacher is often expected to demonstrate in order to function effectively in and out of the language classroom. Chapter 3 also looks at needs analysis as a means of collecting empirical data on knowledge and skills required by language teachers in various educational contexts. Chapter 4 studies general principles of language assessment and peculiarities of assessment of language teacher language competence.

3.1. Language teacher language competence: general principles

The concept of FL teacher competence has been a focus of research since the 1980s (e.g. Kennedy, 1983; Thomas, 1987; Wright&Bolitho, 1993; Cullen, 1994; Medgyes, 1999; Medgyes, 2001; Trappes-Lomax, 2002; Richards, 2001; Richards, 2010). Kennedy (1983) was the first to stress that language development of language teachers

required a special approach, different from general language courses. He suggested the division of teacher language needs into ‘course needs’, i.e. those trainee teachers might require during a course of studies at college/university – from listening to lectures to writing essays; and ‘teaching needs’, that ‘reflect the role of the course participant as a teacher and predict the skills that the teacher will need after the course’ (1983: 76). To describe teacher needs, Kennedy introduced the term ‘teaching activities’ – the tasks that a teacher undertakes during a working day which involve the target language in some way (1983: 77). Although such a definition might seem quite vague, it was further specified through a classification of these activities:

- a. selecting and evaluating material;
 - b. preparing lessons;
 - c. supplementing textbook exercises and designing own materials;
 - d. conducting a lesson;
 - e. setting and marking exercises, tests and exams
- (Kennedy, 1983: 77).

Kennedy (1983) was among the first to emphasise that for successful performance of the activities above, a ‘specific variety of language’ was required, but there was little idea of what that language was. The only exception, according to Kennedy, was Classroom Language, that by 1983 had received some attention and was one of the areas quite intensively investigated (e.g. Hughes, 1987; Willis, 1987). Although Kennedy did not specify any grounds for the classification above and based it on theoretical sources without involving any empirical data, this publication can be considered, in some way, groundbreaking. Kennedy raised the issue of ‘teacher language’, that by 1983 had not been discussed, and presupposed that teacher language was different from general language development and, therefore should be treated as a type of ESP activity. By classifying teacher activities Kennedy provided some basis for further description of language teacher competence and what would be later called teacher ‘communication domains’ (Common European Framework of Reference, 2001; Sešek, 2007).

The term ‘teacher competence’ was employed 4 years later by Thomas (1987). Similarly to Kennedy, Thomas investigated command of the target language that a FL teacher should demonstrate. Apart from using the term ‘language competence’ for

describing a FL teacher's desirable command of the language, Thomas was one of the first to consider language awareness as a part of this competence:

'teachers ... should themselves have language competence to a greater degree than that expected of their learners. They should also be competent in teaching of language ... The ability to teach language in turn involves explicit knowledge of the language system and how it operates in communication; this may be called language awareness' (1987: 34)

Thomas singled out 3 components of language teacher competence:

- Competence in language teaching, i.e. **pedagogic competence**
- Explicit knowledge of language system and use – **language awareness**
- Competence in language system and use – **language competence** (1987: 35) which the native speaker has but the non-native speaking teacher needs to develop:
 - 'formal' component (phonological, graphological, syntactical, lexical)
 - contextual/ discourse component
 - stylistic component
 - informational appropriacy (theme and rheme, anaphora, etc.)

(Thomas, 1987: 37)

Thomas stressed that the components described were 'skill-dependent' and were employed through listening, reading, speaking and writing. Thomas stated that **language competence** of a FL teacher was closely connected, or interrelated with **pedagogical competence**. In some way similar to Kennedy, Thomas presented the four constituents of pedagogical competence, all of which require use of the target language: management, teaching, preparation and assessment (1987: 37). The *management* component, according to Thomas, meant, first of all, classroom management and, therefore, involved Classroom language to be used for instruction, establishing rapport, managing equipment and materials. The *teaching* component, as described by Thomas, related directly to the process of teaching and involved teacher's knowledge of and about the target language, teacher's language skills and an ability to impart language on learners. The *preparation* component dealt with lesson planning and materials development and was described as a teacher being 'prepared both mentally and physically in terms of both his teaching strategies and his use of resources' (1987: 37).

Thomas, much earlier than Council of Europe in its seminal Common European Framework of Reference (1996; 2001), raised the issue of levels of competence which

are achieved gradually due to its complexity and amount of knowledge and skills. In 1987, this was the problem not thoroughly discussed, and Thomas was one of the first to come out with his own system of levels for pedagogical and language competence of a FL teacher:

- **Key:** *sine qua non* for a language teacher.
- **Essential:** necessary for good language teaching.
- **Needed:** important for good language teaching.
- **Ideal:** useful for a language teacher.
- **Luxury:** not normally required by a language teacher

(Thomas, 1987: 39)

Description and gradation like the one above seem to bring more questions than answers. First, it would be quite difficult to distinguish between ‘key’, ‘essential’ and ‘needed’ skills, with ‘luxury’ and ‘ideal’ being possibly treated as something non-obligatory. Second, the interpretation of level names could also sound misleading – the borderline between ‘essential’ and ‘needed’, or ‘ideal’ and ‘luxury’ could be very vague and unclear. Last but not least is that Thomas did not provide a single example to illustrate the difference between these levels, which could immediately lead to various interpretations.

Despite some issues that Thomas’ work did not discuss, this publication can be treated as extremely important. Whereas Kennedy (1983) considered teacher language competence as an area of ESP as opposed to general language competence, Thomas suggested that language competence of a FL teacher was wider and more complex in comparison to a general language user, i.e. teacher-specific language complementing command of general language.

The concept of teacher language competence got further development almost 10 years later in work of Wright&Bolitho (1993); Cullen (1994); Widdowson (2002); Wright (2002); Bolitho&Carter (2003). These authors emphasised the importance of teacher language awareness (TLA) as an essential element of teacher language competence that allows a teacher not only to use, but also to teach the target language. TLA was considered as linguistic knowledge required by language learners combined with a knowledge of language teaching principles.

Wright&Bolitho (1993), referring to Edge (1988), viewed a language teacher as a language *user*, a language *analyst* and a language *teacher*, emphasizing that lack of language awareness often manifested itself at classroom level. They defined language awareness as ‘awareness of how language works’ and stated it was crucial for accomplishing various teacher tasks – planning lessons; evaluating, adapting and designing materials, testing and assessing learners (1993: 292). Stating importance of LA, Wright and Bolitho, though, did not give any explanation or illustration of its components – knowledge or skills – which language teachers must develop.

The concept of Language Awareness was indirectly touched upon by Cullen (1994) who was mostly interested in practical issues of incorporating a language component into FL teacher training programmes. Cullen criticized his contemporary teacher development programmes which consisted of ‘fairly predictable sets of component parts’:

- methodology;
 - linguistics (primarily theoretical, aiming at developing language awareness);
 - literature;
 - language improvement, that can be linked to special language teachers need in the classroom
- (Cullen, 1994: 162)

Cullen’s major interest lay mostly in enlarging the amount of the target language in teacher training programmes, i.e. conducting other course elements (methodology, linguistics) through English, giving trainees more opportunity to read, listen and discuss relevant things in English. Cullen did not aim at describing language teacher competence or suggesting its model, neither did he specify if any of the other components he referred to needed improvement/reshaping. However, this publication can be considered important for this research because it gave an outline of what a FL teacher should be able to do and how a language course could be organized. Cullen’s understanding of language awareness, which he treated as understanding of how language operates. Similarly to Thomas (1987), Cullen saw language awareness as an important element of teacher language competence. He saw theoretical linguistic courses that teachers take at university as a means of developing language awareness. However, there is no indication if Cullen saw any difference between language

awareness (knowledge of language system and use) and *teacher* language awareness that results in an ability to teach language.

In a similar vein, Medgyes (1999), when speculating on non-native speaker teacher language competence, singled out 3 components of ‘teacher expertise’ – language proficiency (listening, speaking, reading, writing); language awareness and pedagogic skills (1999: 54). Stressing their importance, the author did not make it clear whether these components were/should be developed within general and/or teacher-specific domain, nor provided a detailed description of the constituents.

The concept of teacher language awareness, as opposed to language awareness, was suggested by Wright (2002) who stated that successful language teaching required proficiency in language use, knowledge of language and knowledge of teaching methods which were interdependent and resulted in linguistic and pedagogic sensitivity to the problems of students. One of Wright’s key ideas was combination of language development courses for teachers with courses on ELT methods or pedagogy, where trainees would examine learner language, analyse classroom talk, examine teaching materials to see how linguistic content is handled and have ‘language improvement’ tasks for themselves:

‘Language awareness ... is a way of ... bringing about a closer relationship between content knowledge and classroom methodology. Language education practitioners are involved not in language and teaching but in language teaching’ (Wright, 2002:115)

Widdowson (2002), when speculating on language teacher competence, specified teacher language awareness – explicit knowledge about language and how it works, including language use in the classroom (2002: 105). Widdowson opposed it to language awareness:

‘my ... point is that knowing a language as a subject is not the same as knowing it as it naturally occurs in the social contexts of everyday life’ (2002: 68).

According to Widdowson, teacher language awareness makes a language learnable, a language which has been pedagogically treated so that it is made ‘less alien’ and more accessible to learners (2002: 78).

Developing Widdowson’s idea, Trappes-Lomax (2002) saw as essential dimension of professional development shifting emphasis from teachers ‘thinking about the language to thinking about the practical side of working with the language for teaching

purposes' (2002: 8). Similarly to Widdowson, Trappes-Lomax thought of involving both communicative proficiency and consciousness of language into language teacher education programmes. This idea was further developed by Bolitho (2003), when he speculated why language awareness 'remained on the periphery of ...language teacher education' (2003: 251). Supporting the model of pre-service language teacher education based on 'language systems' component (grammar, phonology, semantics) and 'language improvement' component, the author nevertheless was concerned about their sufficiency for FL teacher's job.

'...neither proficiency in a language nor knowledge about that language are sufficient on their own to equip a teacher to teach it. Trainee teachers need to be able to analyse language, to apply different strategies for thinking about language (analysing, contrasting, structuring) in order to be able to plan lessons, to predict learners' difficulties, to answer their questions, and to write and evaluate materials' (Bolitho, 2003: 255)

Paying much attention to the crucial role of TLA for teacher effective functioning, Bolitho emphasised the paradox of the situation, i.e. a most important dimension of FL teacher development getting least attention at the pre-service level.

Whereas the publications reviewed above mostly dealt with language awareness that FL teachers required, a very important step in describing FL teacher language competence was taken by Richards (2010). In contrast to Wright (2002), Widdowson (2002) and Bolitho (2003), Richards did not investigate teacher language awareness, but suggested a list of communicative skills a language teacher was expected to demonstrate in the target language:

- to comprehend texts accurately;
- to provide good language models;
- to maintain use of the target language in the classroom;
- to give explanations and instructions in the target language;
- to provide examples of words and grammatical structures, give accurate explanations (for example, for vocabulary and language points);
- to use appropriate classroom language;
- to select target language resources (e.g. newspapers, magazines, internet);
- to monitor his/her own speech and writing for accuracy;
- to give correct feedback on learner language;
- to provide input at an appropriate level of difficulty;

- to provide language enrichment experience for learners.

(Richards, 2010: 3)

This represented the first attempt of a language skill taxonomy which would demonstrate that FL teacher communicative skills differ from communicative skills of other language users. This list can be considered as a good basis for further development of a teacher language skill typology with a few things classified and explained. For example, for text comprehension Richards' work provided no indication on text types, topics, length and whether the texts should be written (reading) and/or oral (listening). Apart from these, there was some overlap in skills description, for example, '*to maintain use of the target language in the classroom*' sounds similar to '*providing language enrichment experience for learners*'. Although some skills singled out by Richards were defined quite vaguely, they all seem essential for teacher job. Many of them could be a good illustration for teacher language awareness, e.g. providing good language models, giving explanations in the target language, providing input at an appropriate level of difficulty. Some skills can be classified as Classroom Language skills, for example giving correct feedback or using appropriate language means. Richards' work can be considered seminal for this research, as it provided some substantial ground for further skills description, language awareness and requirements to FL teachers in general. The taxonomy of skills suggested by Richards serves a basis for the needs analysis in this research (see part 3.3 and Chapter 6).

Two more publications, quite different in nature and purpose from the publications above, are viewed as essential for his research. Spratt (1996) and Thornbury (1997) in their language books for teachers presented through a series of tasks what a language teacher was supposed to know in terms of grammar, phonology, vocabulary of the target language. Neither of the authors referred to any syllabus or curriculum for language teacher development to provide a background for content selection, although both Spratt and Thornbury expressed some criticism of their contemporary programmes for language teacher training. According to Thornbury, a major weakness of both in-service and pre-service courses was lack of emphasis on TLA that he defined as 'knowledge that teachers have of the underlying systems of the language that enables them to teach effectively' (1997: x). Apart from the definition of teacher

language awareness, both books provide a series of tasks that aim at building this awareness, i.e. at combining knowledge about language with pedagogical knowledge. These tasks, according to Thornbury, can be seen as 2 groups: *language tasks* (e.g. matching, categorisation) and *pedagogical tasks* (e.g. giving feedback on pupils' error; evaluating teaching materials).

According to Thornbury, lack or absence of TLA can have serious consequences in terms of poor tuition standards:

- a failure to anticipate learners' problems;
- an inability to plan lessons at the right level;
- an inability to interpret syllabuses and materials and to adapt them to the specific needs of the learners;
- an inability to deal with errors or to answer learners' questions;
- a general failure to earn the confidence of the learners due to a lack of basic terminology and ability to present new language clearly and efficiently.

(Thornbury, 1997: xi-xii)

Spratt's (1996) and Thornbury's (1997) publications are viewed as important for this research because not only do they present Teacher Language Awareness (TLA) as a key component of teacher language competence, but also give an insight into TLA by providing its definition and describing the ways it can be developed throughout teacher training courses.

Further literature search demonstrated only isolated fragmentary opinions about aims of language teacher development programmes from 'preparing teachers for linguistic emergency' (Marton, 1988: 99) to 'improving teacher proficiency in the language either generally or with specific pedagogic purposes in mind' (Berry, 1990) and 'developing study skills and skills that the teacher will need after the course' (Kennedy, 1983: 76). There is another group of publications that is reviewed separately, although it can be referred to as published articles, too. This group is **reports on research projects** performed at different institutions and within different time scales – from several weeks to several years.

3.2. Language teacher language competence in national and international research projects

All represented projects (1994-2007) aimed at gaining a better understanding of what a teacher of modern languages was supposed to do/know. Some projects (e.g. Language Proficiency Assessment for Teachers of English (LPATE), Hong Kong) went further than description of teacher language skills and investigated how various areas of teacher knowledge and skills can/should be evaluated.

The Singapore '**English for teaching purposes**' project (1994) looked at what language means were required by primary teachers to carry out various tasks, aside from the classroom context. The aim of the project was 'to make sure they [teachers] can cope with language uses arising in the teaching contexts' (Skuja&Mee, 1994: 163). The data was collected through a series of brainstorming sessions with teachers of English in Singapore, and as a result of those sessions 'a list of all possible tasks for beginning teachers was drawn up' (1994: 162). Those were mostly situations found in schools: apart from classroom interaction, which was intentionally left out, the situations included assembly, communication in staffroom, and parent evenings. As the project outcome, the researchers came out with three large areas of teacher language use:

- language teachers need for getting information
- language for teaching
- language appreciation. (Skuja &Mee,1994: 165)

Whilst the first two areas seem to be quite transparent, although not defined in detail, the third one, 'language for appreciation', would definitely benefit from further explanation. The article under review focused mostly on the results achieved and conclusions made, whereas methods of data collection were not specified. Observation was once mentioned, but there were neither samples of observation framework presented, nor methods of data analysis described. Although this research project involved quite limited research population - 26 in-service primary school teachers (1994: 164), it suggested the components and possible communication areas that comprise teacher language use.

The project of the Hong Kong Polytechnic – **Language Proficiency Assessment for Teachers of English** (2000) – aimed, first of all, at defining the communicative skills an English teacher needed in and out of the classroom and, second, at designing a new format of examination for English language teachers (in-service level). The first stage of the project was a taxonomy of skills that language (English) teachers required for effective functioning. This stage involved a substantial amount of observation and a lesser amount of teacher interviews that helped to identify the communicative skills teachers needed (Coniam&Falvey, 2002). As a result, 4 groups of skills – listening, reading, speaking and writing - were specified with concrete examples of what the respondent teachers did most frequently in and out of the language classroom. Reading was specified as mostly reading articles on professional matter; writing included writing letters, error correction and feedback on student written work. Speaking, apart from Classroom English, included reading aloud (poetry and prose) and storytelling. Language awareness, although it was not singled out as a separate skill/knowledge area, was treated as a prerequisite for many skills above. For example, error correction that was viewed as one of writing skills, was interrelated with language awareness; a similar situation could be observed with speaking.

Once the key communication skills were identified and prioritised, the second stage of the project was implemented: design of a language examination for language teachers. The examination was assessment of language ability and not pedagogy or teaching methodology (Coniam, 2013). The examination was occupation-based which can be traced through both its content and format¹⁶, and assessed all four communicative skills and Classroom English that was treated as a part of speaking ability. The research report contained neither information on the further use of the examination, nor empirical data from its piloting. Nevertheless, it can be viewed as crucial for this study because it singled out the communicative skills that the Hong Kong teachers needed, and suggested some tasks and techniques for assessment of those skills.

European Profile for Language Teacher Education: a frame of reference (2004) was designed to promote a profile for language teacher education in the 21st century,

¹⁶ The format and content of the LPATE examination is reviewed in Chapter 4

following the profile for general language learning – Common European Framework of Reference (2001). As its authors – Kelly and Grenfell – stated in the Introduction,

‘It [the document] aims to serve as a checklist for existing teacher education programmes and a guideline for those still being developed’ (2004: 4)

The document proposed that Foreign Language teacher education should include the following elements:

1. Training in language teaching methodologies ... and classroom techniques and activities.
2. Initial teacher education that includes a course in language proficiency and assesses trainees’ linguistic competence¹⁷.
3. Training in information and communication technology for pedagogical use in the classroom.

Referring to the model of competence employed in the Common European Framework of Reference (2001), the authors admitted it could not at that moment be a commonly accepted model, as the Common Framework ‘has not been adopted in equal measure throughout Europe’ (Kelly&Grenfell, 2004: 38). Nevertheless, empirical data collection that involved about 30 universities across the EU, was based on the Common Framework competence model. The data was collected mainly through surveys, when the participants were offered closed- and open-ended questionnaires. The research aimed to suggest a detailed taxonomy of sub-skills for listening, reading, speaking and writing that teachers of modern foreign languages require and universities develop with their trainees. However, according to Kelly and Grenfell (2004), the study seemed to have faced a problem similar to the one described in the Common European Framework: the higher the level of language competence is, the more difficult it is to describe, although it provides a greater chance to be creative and confident. As a possible consequence, the project suggested some broad descriptions of communicative skills that European universities develop with their trainees, for example, ‘improving linguistic competence and achieving near native competence in the target language’ (Kelly&Grenfell, 2004: 50).

¹⁷ ‘Linguistic competence’ is used interchangeably with ‘language competence’ meaning the ability to use the target language in different situations, that becomes clear from the sentence ‘All EU citizens should have linguistic competence in their own mother tongue and 2 other languages’ (Kelly & Grenfell, 2004: 11)

Standards for Teachers of Indonesian Project¹⁸ (Australian Federation of Modern Language Teachers Associations, 2005), saw the key element of content knowledge for language teaching as a knowledge of the language being taught (2005: 14). The authors emphasised that

‘the highly contexted nature of language use in language teaching, i.e. language proficiency for language teachers is not a simple question of their measurement on the scale of proficiency, but rather their ability to use their language to enact language pedagogy. Language proficiency cannot be fully understood outside the teaching context in which the proficiency is used’ (2005: 14)

With a reference to Wright and Bolitho (2001), the document stated that the definition of content knowledge of a language teacher was more than a question of proficiency of a general language user, because it should include additional knowledge about the language (2005: 14). Among the types of knowledge a language teacher must acquire, ‘knowledge of the language’ and ‘knowledge about the language and language in general’ were the first listed. Other positions were occupied by knowledge about the target culture and pedagogical content knowledge.

The document clearly outlines the following areas of teacher language that are considered essential in and out of the language classroom:

- classroom language, including classroom management;
- communicative skills of reading and listening, speaking and writing;
- language awareness that manifests through an ability to teach the target language using relevant vocabulary, syntax and structure (Australian Federation of Modern Language Teachers Associations, 2005: 55).

The **project of the University of Split** (2005) was built around competences that a future language teacher should develop in the undergraduate teacher training programme. The students of the English language and literature department were asked to write an essay in which they described the competences a graduate should develop. They were expected to write about knowledge, abilities and/or skills that would help them to teach English effectively. The essays were analysed, the competences listed

¹⁸ Although the project does not deal with English as a foreign language and English language teacher preparation it provides very useful data for a foreign language teacher’s language competence description. Indonesian is considered a foreign language in Australia which makes the project results relevant for this research

and ranked by students in the order of importance. The most important one was *communicative competence*, i.e. the ability to use the target language accurately, appropriately and fluently in different speech situations (Ćurković-Kalebić, 2005: 110). The second competence, closely connected with the first one, is *communication and presentation skills* – the teacher’s ability to speak clearly, to be interesting to learners and to be understood by them. The major critique of this small-scale research might be some lack of reliability of results, with the research population being quite limited and only one research method applied. Nevertheless, its findings are in some good keeping with other research projects and emphasise the importance of speaking skills in FL teachers’ professional life.

One more important contribution was made by **the University of Ljubljana** (Sešek, 2007) within the project “*English for teachers of English as a foreign language – toward a holistic description*”. The project carried out in 2003-2005 aimed at needs analysis of Slovenian teachers of English as a Foreign Language in and outside of the language classroom. The needs analysis was supposed to contribute into revision of teacher training aims by compiling a list of activities and competences that teachers of English in Slovenia need to develop.

The initial stage of the project was literature review that revealed that such studies were rare, and the bigger part of English language teacher training and development, at least in Europe, was still

‘nested within traditional language and literature studies, where the students’ target language proficiency development is often marginalized as a curricular goal and conceptualized as English for General Purposes rather than profiled from the point of view of the graduates’ future profession’ (Sešek, 2007: 412).

Sešek (2007) supported the idea of Kennedy (1983) and Elder (2001) that teacher English was a case of ESP, although not entirely comparable to all ESP types. This vision of teacher language competence provided some ground to look critically at the existing FL teacher training models which usually aim at developing trainees’ general language competence and overlook a teacher-specific component (Sešek, 2007: 412). Sešek referred to several ways of describing teacher language competence – from description of teacher talk (Chaudron, 1988) and lexico-grammatical descriptions of teacher language (Hughes, 1981; Spratt, 1997) to more specific frameworks such as

‘Language Proficiency Assessment for Teachers of English’¹⁹ (2000) or ‘A Language Profile for a FL Primary Language Teacher’ (2002). Having compared various ways of describing teacher language competence, Sešek (2007) gave preference to the general language competence model (Common European Framework, 2001), that served a framework for the planned analysis of language need of teachers of English in Slovenia.

Sešek’s research (2007) involved more than 100 participants in interviews, classroom observation and teacher diaries. The data collection aimed to investigate communicative tasks that teachers perform, and teachers’ communicative language competences (2007: 414). The data from all the respondents was summarized and ‘the relative significance of each skill was established’ (2007: 419). The criterion for defining skill significance was the amount of time per week the respondents were involved in it. As a result, the major outcome of the project was a taxonomy of communicative skills that FL teachers in Slovenia require for their job. Speaking skills were on top of the list, followed by reading, writing and listening. The taxonomy of skills served a dual purpose:

1. The skills that Slovenian teachers employed were compared to the general language skills presented in Common European Framework (2001). This resulted in a more detailed skill description for both general and professional purposes.
2. The taxonomy of skills compiled gave the researchers some ground to suggest changes in language teacher development programmes, i.e. changing the focus of language courses.

The outcomes of the Slovenian project (Sešek, 2007) are considered valuable for this research because the project specified some elements of language teacher language competence – speaking and Classroom English, reading, listening and writing, and also vocabulary and grammar. The importance of all 4 communicative skills was highlighted, although some skills (e.g. classroom language) are required more often than others (e.g. listening).

Although the projects under study (Singapore, Australia, Hong Kong, Croatia, Slovenia) differed in terms of time scale and methods of data collection, they all

¹⁹ Referred to earlier in this chapter (pp.36-37)

contributed to understanding of the structure of language teacher language competence. The major outcomes of the 6 projects are presented in Table 3.1.

Table 3.1. Summary of project findings on language teacher language competence

	Methods	Project population	Focus	Outcomes
Singapore 1994	-survey -observation	primary school FL teachers of English (in-service level)	the language a teacher needs outside the classroom	3 areas singled out: - language for information - language for teaching - language appreciation
Hong Kong 2000-2002	- lesson observation - teacher interviews	school teachers of English	communicative skills in the target language a teacher needs	A new format of examination for English language teachers (reading, speaking, writing, listening)
European Profile 2004	-survey -literature and document review	staff of European colleges and universities involved in teacher development	a FL teacher's skills (pre-service level)	a framework for teacher development curriculum design and evaluation
Croatia 2005	-essay analysis -ranking	students of English language and literature (pre- service level)	professional skills a FL teacher needs	a list of 14 competences compiled and ranked with communicative competence on its top
Australia 2005	-literature review -teacher interviews -survey	teachers of Indonesian as a FL in Australia (in- service level)	7 types of knowledge a teacher should acquire	a FL teacher's communicative competence described - general language - classroom-related language
Slovenia 2007	-literature review -interview -lesson observation -teacher diaries	beginner and experienced English teachers in Slovenia (in-service level)	Competences teachers require in and outside the language classroom	A list of activities (ranked in order of frequency) teachers are involved in; recommendations for teacher development programme improvement

The most important contribution to this research is seen in the Slovenian project. First, the context of the project is quite similar to this research context – training of non-native English speaker teachers in a similar environment. Although the Slovenian project focuses on in-service teacher development whilst the current research deals with pre-service training, its outcomes are considered relevant and essential for:

- defining the focus of FL teacher training and assessment;

- designing data collection instruments for empirical data on language needs of English teachers in Russia (Chapter 6).

As can be seen from Table 3.1, all reviewed projects came to the conclusion that a language teacher language competence comprises the 4 key communicative skills. Speaking is viewed as one of the prevailing skills (Singapore, 1994; Croatia, 2005; Slovenia, 2007). Classroom Language is seen as a separate skill (e.g. Singapore, 2000) and sometimes is opposed to general speaking (e.g. Australia, 2005). Other projects (e.g. Hong Kong, 2000; Slovenia, 2007) see Classroom Language as an inseparable part of teacher speaking skills. All represented projects see reading and writing as the skills frequently required by language teachers, with reading on ELT issues and writing lesson plans being most frequent activities (e.g. Slovenia, 2007). The Hong Kong project also considers reading aloud as a key skill for language teachers (Coniam, 2013: 150). The majority of projects see listening as an important skill for language teachers (e.g. Hong Kong, 2000; Australia, 2005; Slovenia, 2007), although some data demonstrated that it is the skill less frequently employed by teachers (Sešek, 2007).

3.3 Towards a working definition of foreign language (FL) teacher language competence

The review of publications in the field of language teacher development - project reports, documents and exam syllabi demonstrated some similarities in how language competence of a language teacher is viewed in different countries (part 3.2). However, no definition of language teacher language competence has yet been suggested. Some authors (e.g. Sešek, 2007) and documents (e.g. Language Education Policy Profile for Poland, 2005; Requirements to Bulgarian teachers of modern foreign languages, 2006; European Profile for language teacher education, 2004) refer to the concept of general language competence adopted by the Common European Framework of Reference (2001). According to the Common European Framework of Reference (hereafter CEFR), language competence is seen as a set of interrelated elements:

- linguistic: not only the range and quality of knowledge – lexical, phonological, syntactical, but also cognitive organisation and the way this knowledge is stored (Common European Framework, 2001: 13);
- sociolinguistic: knowing how to use and respond to language appropriately, given the setting, the topic, and the relationships among people communicating (www.nclrc.org; retrieved on December 2, 2012);
- pragmatic: functional use of linguistic resources; the mastery of discourse, coherence and cohesion, the identification of text types and forms, irony and parody (Common European Framework, 2001: 13).

Language competence is seen as a key pre-requisite for language users to engage in various language activities. Those are seen as understanding (listening and reading), speaking (interaction and production) and writing (2001: 9).

Whilst some agreement has been reached about general language competence, language competence of a foreign language teacher is still seen as a number of elements that complement general language competence. Almost all authors (e.g. Thomas, 1987; Cullen, 1994; Medgyes, 1994; Spratt, 1996; Thornbury, 1997; Widdowson, 2002; Wright&Bolitho, 1993; Wright, 2002) refer to *language awareness*, or teacher language awareness, as an essential element of teacher language performance in and out of the classroom. Teacher language awareness can be seen as a part of linguistic competence or as a separate element that allows to *teach* a language. While linguistic competence, according to CEFR is expected to be demonstrated by any language user, teacher language awareness is a unique area of foreign language teacher knowledge. In this research, *linguistic competence* is treated as knowledge of phonological, grammatical and lexical systems of language and the target language (English), an awareness of how language works (Common European Framework of Reference, 2001). Teacher language awareness, is understood as knowledge of phonology, grammar and vocabulary of the target language and the ways they function in order to teach the target language effectively, i.e.:

- deal with errors effectively; anticipate possible problems;
- plan lessons and design materials at the right level;

- present and practice language items effectively. (based on: Bolitho, 1993; Widdowson, 2002; Wright, 2002, Bolitho&Carter, 2003).

The majority of publications referred to in parts 3.1-3.2 raise the question of language activities that a FL teacher is involved in. Since the 1980s, authors have been emphasizing the importance of activities in all 4 skills – listening, reading, speaking and writing (Thomas, 1987; Cullen, 1994; Medgyes, 1994; Spratt, 1996; Ćurković-Kalebić, 2005; Sešek, 2007; Richards, 2010). At the same time, only a few publications (e.g. Ćurković-Kalebić, 2005; Sešek, 2007) make an attempt to specify these activities and/or sub-skills they require by providing an outline of content areas, task/text types that teachers are expected to deal with. Classification of language activities seems quite a complicated task due to an integrated character of those activities, i.e. more than one skill and/or knowledge area being involved. Richards' (2010) contribution (Chapter 3, p.33-34) is considered important for this research as one of the first attempts to classify teacher language activities and sub-skills they involve. Emphasising the importance of all four communicative skills for a language teacher Richards, however, concentrated mostly on classroom language and teacher language awareness – from providing good language models to giving correct feedback (2010: 3).

No clear distinction has been observed in literature between general and teacher-specific skills. It can be presumed that the difference lies in communicative domains (Common European Framework of Reference, 2001), text and task types that teachers deal with. Whilst general communicative skills have been described in detail in literature and documents, teacher skills got more limited attention (e.g. Ćurković-Kalebić, 2003; Sešek, 2007). For this research, the key point is that all four communicative skills – listening, speaking, reading and writing – need to be developed.

Language needs of and requirements to language teachers may vary significantly from country to country. Moreover, within each country a difference in needs can be observed. Although no analysis of teacher needs has been performed in Russia, it can be presumed that language needs of language teachers in big cities (e.g. Moscow, St Petersburg, Nizhny Novgorod) might differ from those of teachers in small towns and villages. The amount of knowledge and skills that language teachers are expected to

demonstrate depend on various factors – from the level they work at to a general context of teaching (e.g. teaching aims; resources available; qualification requirements to FL teachers; availability of a target language outside the classroom, etc.).

Jasso-Aguilar (2015) suggested considering the following dimensions when defining learner language needs and, consequently, aims of training:

- general personal background of learners (trainees);
- language background;
- attitudinal and motivational factors.

Although the importance of needs analysis for curriculum planning is obvious, and a substantial number of needs analyses have been published (e.g. Vandermeeren, 2015; Gilabert, 2015; Cahudron, Doughty & Kim, 2015), there has been quite a little research on needs analysis itself.

Hutchinson and Waters (1990) wrote about analysing a target language situation that includes ‘necessities’, ‘lacks’ and ‘wants’ of a language user (1990: 55-58). According to Hutchinson and Waters, analysis of a target language situation should include:

- purpose of language use (study, work, training, promotion, etc.);
- medium of language use (spoken/written; telephone/face-to-face) and type of discourse (e.g. academic texts; informal conversation; technical manuals);
- content areas of use and level of use;
- type of interlocutor (native/non-native speaker; colleague/teacher/friend, etc.);
- setting of language use (home country/abroad; office; lecture theatre, etc.).

Quite similarly to Hutchinson and Waters (1990), Yalden (1995) when writing about needs survey suggested getting information on learner general background, language needs and learning styles and preferences (1995: 130-131).

Richards (1997), referring to Munby (1978), suggested obtaining information on:

- the situations in which a language will be used;
- the purposes for which the language is needed;
- the types of communication that will be used (written/spoken; formal/informal);
- the level of proficiency that will be required (1997: 243).

Jasso-Aguilar (2015) and Long (2015) considered analysis of tasks that language learners are supposed to deal with as a major purpose of needs analysis. According to Jasso-Aguilar, analysis of tasks would result in description of skills that learners need to develop; content areas and purpose of language use, as well as type of language required for task fulfilment.

For this research, analysis of language teacher needs in Russia is seen as a key step in defining a focus of language development and, consequently, assessment. Language teacher needs analysis aims to specify the areas of language teacher competence as seen in theoretical and empirical research. Thus, analysis of language needs of English teachers is designed to specify:

- areas of classroom language that teachers require;
- language activities in listening reading, speaking and writing that teachers are involved in their everyday teaching practices;
- areas of teacher language where teachers feel confident.

Procedures and implementation stages of the language teacher needs analysis in Tula region (Russia) are presented in Chapter 6: Research Methodology.

Chapter 4

Literature Review Part II: Language Testing for Language Teachers: national and international experience

Publications in the area of language assessment for pre- and in-service language teachers revealed a gap between an abundant range of research on general language testing – books, articles and research reports, and quite a limited number of articles on language testing/assessment of language teachers. Theoretical and empirical researchers and examination bodies in different countries give attention to general issues of language testing – types and forms of tests, test design; providing validity and reliability of language tests and marking procedures, rater training, etc. At the same time, language tests/examinations for teachers of foreign languages receive much more limited attention of researchers.

Chapter 4 starts with consideration of key dimensions of language test validation: validity, reliability, authenticity and practicality. Part 4.2. takes an insight into the difference between direct and indirect testing that seems important in designing language tests for teachers. Part 4.3 discusses advantages and disadvantages of various task types and their applicability to valid and reliable assessment of basic communicative skills – listening, speaking, reading and writing. The selection of sources on the issues of general language testing is more selective than exhaustive because of an overwhelming amount of publications in this area. Only seminal publications on language testing are studied here, such as Alderson 1995; Heaton, 1995; J.D.Brown, 2000; Hughes, 2003; H.D.Brown, 2004; McNamara, 1997; McNamara, 2006; Bachman&Palmer, 2010.

Then the chapter narrows down to the publications on language examinations for language teachers (Consolo, 2008; Elder, 2000; 2001; Grant, 1997). The major aim here (part 4.4) is to examine major differences between ‘general’ and ‘teacher’ language testing. Part 4.4. also attempts to single out possibilities and constraints that language testing for language teachers has to face, in comparison to general language testing.

Part 4.5 focuses on practical implementations of theoretical findings on language testing for language teachers and reviews the existing national and international

language examinations for teachers of foreign languages. Chapter 4 concludes by defining some key objectives in language assessment for language teachers and ways these objectives can be achieved through examination format and content.

4.1. Language test evaluation: major dimensions

Study of fundamental publications on language testing (Alderson, 1995; Heaton, 1995; McNamara, 1997; Brown J.D., 2000; Hughes, 2003; Brown H.D, 2004; Norris, 2009) demonstrated some similarity of opinions on what requirements a language test should meet. Alderson (1995) concentrated mostly on validity that can fall into several categories, and reliability. Similarly to Alderson, Heaton (1995) paid much attention to validity and reliability of tests ‘whether it be a short, informal classroom test or a public examination’ (1995: 159). H.D.Brown (2004), speculating on principles of language assessment, identified 5 criteria for test validation:

‘How do you know if a test is effective? For the most part, that question can be answered by responding to such questions as: Can it be given within appropriate administrative constraints? Is it dependable? Does it accurately measure what you want to measure? These and other questions help to identify five cardinal criteria for “testing a test”: practicality, reliability, validity, authenticity and washback’ (2004: 19).

Richards (1997), when speculating on major test characteristics, wrote about test validity and its types, and also test reliability (1997: 314, 396). MacNamara (1997), discussing performance language tests, concentrated on several types of validity, with reliability being mentioned as an essential test characteristic but not getting the author’s detailed attention. J.D.Brown (2000) mostly dealt with validity issues of language tests, although he also considered reliability as a key requirement to any test.

Further study demonstrated that the categories of validity and reliability are treated by various authors as commonly accepted. Researchers seem to focus on various dimensions of validity and reliability of language tests, whilst it is hard to pinpoint a consensus on definitions to these terms. Chapelle (2011) investigated common issues of validity and ‘ways of arguing it (2011: 19). Xi (2010), Kane (2010), Davies (2010) took an insight into test fairness. They treated it as an aspect of validity and sometimes called it ‘comparable validity’ (Xi, 2010: 147) for all relevant groups of test takers. Davies (2011) treated validity as an essential test characteristic and discussed ways to achieve it.

Reliability is often viewed as reliability of test items (Alderson, 1995; Heaton, 1995; H.D.Brown, 2004) and reliability of test administration that includes rater reliability in general (Bachman&Palmer, 2010; Kuiken&Verde, 2014) and rater behaviour in particular (Lumley, 2002; Fulcher & Davidson, 2011; Ling, Mollaun, Xi, 2014). Both theoretical and empirical research conducted recently demonstrated that rating issues cause the majority of construct-irrelevant factors (such as improper use of rating scales, disagreement between raters, rater fatigue) and affect both test validity and reliability. For this research, rater reliability is seen as crucial due to the open-ended task format of the current Final Language Examination under study that makes Exam takers' expected performance difficult to describe and, therefore, to mark.

Authenticity of test tasks is considered an important issue by Bachman (1996, 2010), McNamara (1997), Lewkowitz (2000), Hughes (2003), H.D.Brown (2004) and is usually seen as 'imitation of' or similarity of test tasks to life circumstances. Provision of authenticity through choice of test tasks and test formats may cause threats to test practicality, especially in case of direct tests (part 4.1.3).

East (2015), speculating on language test evaluation, suggested, in addition to considering its validity, reliability, and authenticity, that the following dimensions should also be addressed:

- test *interactiveness*, i.e. whether students can engage meaningfully with the assessment task
- test impact, i.e. whether the assessment leads to positive consequences for those being assessed, for example, to comparatively less stress than a different kind of assessment²⁰ (2015: 106-107)

Internet and computer testing added some specific features to language tests (Dunkel, 1999; Barkaoui, 2014) but did not change the vision of major test parameters. Although computer and internet testing seem to be developing rapidly²¹ and acquiring both supporters and skeptics, major concerns still lie within validity, reliability and practicality issues. While some researchers are concerned about appropriacy of

²⁰ By some researchers (e.g. Alderson, 1995; Heaton, 1995; Brown, 2004) the test impact factor, as described by East, is treated as 'washback effect', or 'beneficial backwash' (Hughes, 2003)

²¹ Many language examinations are available in both paper-and-pencil and computer/internet form, with some of them (e.g. TOEFL, Praxis) existing only in the format of iBT (internet-based testing)

computerized tests for assessing particular skills, others think of construct-irrelevant variables such as computer familiarity or computer anxiety.

In this research, the major parameters of language tests/examinations, and, therefore, test evaluation are seen as:

- validity (part 4.1.1)
- reliability (part 4.1.2)
- authenticity and practicality (part 4.1.3).

4.1.1. Language test validity

The concept that involves quite different interpretations and opinions is *validity*. As H.D.Brown claimed, ‘there is no final absolute measure of validity, but several different kinds of evidence may be invoked in support’ (2004: 22). Chapelle (2011) speculating on issues of providing language test validity, stated that although much has been published on validity of tests, ‘arguing validity for real test interpretations and uses has found the guidance anything but simple’ (2011: 19).

Defining the concept of language test validity, most authors single out different validity types. Alderson (1995) saw validity as ‘testing what the test is supposed to test’ (1995: 170) and distinguished between rational (content), face, empirical, construct and concurrent validity (1995: 171). Richards (1997) described validity as ‘the degree to which a test measures what it is supposed to measure, or can be used successfully for the purposes for which it is intended’ (1997: 396). He suggested division into content, construct, criterion-related validity, and also empirical, predictive and face validity (1997: 396).

J.D.Brown (2000) defined validity as ‘the degree to which a test measures what it claims, or purports, to be measuring’ (2000: 8). He considered 4 types of validity – content, criterion-related (concurrent), construct and predictive and suggested treating them all as ‘different facets of a single unified form of construct validity’ (2000: 8). Similarly to Brown (2000) and others, Hughes (2003) distinguished between content, criterion-related, concurrent validity, validity of scoring and face validity (2003: 33-34). H.D.Brown (2004) saw validity as the most complex criterion of an effective test:

‘How is validity of a test established? There is no final, absolute measure of validity, but several different kinds of evidence may be invoked in support. In some cases, it may be appropriate to examine the extent to which a test calls for performance that matches the unit of study being tested. In other cases, we may be concerned with how well a test determines whether or not students have reached an established set of goals... (2004: 22).

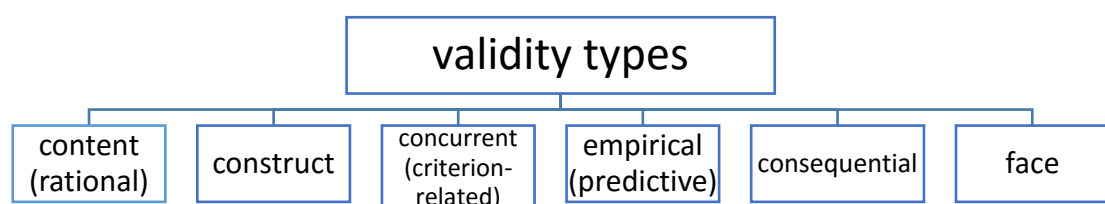
Another way of measuring test validity, according to H.D.Brown, is correlation with other related but independent measures.

As can be seen, most of the reviewed authors subdivided validity of language tests into content, construct, concurrent, predictive and face validity. Research conducted within last several years (2010-2015) did not introduce any new types of validity, but rather concentrated on ways of providing (arguing) validity. Davies (2011), referring to Anastasi (1988) emphasised that validity of a test cannot be reported in general or abstract terms, and no test can be said to have high or low validity in the abstract. According to Davies, validity can be established only with reference to a particular use for which the test is designed:

‘In language testing validity cannot be achieved directly but only through a process of validation: we validate a test and then argue that it is valid. The analogy to the relation between justice and the law is apt: justice is not attainable directly and has to be reached for by way of the law’ (2011: 38).

Figure 4.1 summarises the major types of language test validity that were addressed by different authors (Alderson, 1995; Heaton, 1995; Richards, 1997; J.D.Brown, 2000; H.D.Brown, 2004; McNamara, 2006).

Figure 4.1. Major types of language test validity



Content validity has been in the focus of attention of many authors, both in its theoretical and empirical dimensions. Alderson (1995) defined content validity as ‘representativeness or sampling adequacy of the content [of a test] (1995: 172). According to Alderson, content validity depends on logical analysis of a test content to see if the test contains a representative sample of the relevant language skills.

In a similar vein, Hughes (2003) and H.D.Brown (2004) emphasised the importance of test content constituting a representative sample of language skills, structures, etc. which the test is meant to assess. J.D.Brown (2000) also considered content validity as the degree to which a test was a representative sample of the content that the test was originally designed to measure (2000: 8).

Construct validity is treated as ‘extent to which the test is successfully based on its underlying theory’ (Alderson, 1995: 172), while the theory itself is not called into question, it is taken for granted. According to Alderson, the issue is ‘whether the test is a successful operationalization of the theory’ (1995: 182). Heaton (1995) defined test construct validity as capability of measuring certain specific characteristics in accordance with a theory of language behaviour and learning:

‘This type of validity assumes the existence of certain learning theories or constructs underlying the acquisition of abilities and skills. (...) if a communicative approach to language teaching and learning has been adopted throughout a course, a test comprising chiefly multiple-choice items will lack construct validity’ (1995: 161).

H.D.Brown (2004) also addressed construct validity as whether ‘the test actually taps into theoretical construct’ (2004: 25). Treated in this way, construct validity may be viewed as interrelated with content validity (J.D.Brown, 2000: 12). This can be supported by the opinion of McNamara (1997) who, referring to Weir (1988) wrote about some overlap between content and construct validity because ‘we need to talk of the communicative construct in descriptive terms, and as a result, we become involved in questions of content relevance and content coverage’ (1997: 18).

The complexity of the issue was stressed by Kane (2012) who treated construct validity as overwhelming in its scope and therefore largely unfeasible for most practitioners. He seemed to be critical of the existing uniform model of defining test construct validity although did not suggest any alternatives. Later, Kane wrote that ‘the kinds of validity evidence that are most relevant are those that support the main inferences and assumptions in the interpretive argument, particularly those that are most problematic’ (2012: 10). In some way similarly to Kane, Norris (2009) saw construct validity of language tests as an entity of several components. He claimed that construct validity should be measured at the level of test content and the level of test scores, i.e. what

interpretations are made on the basis of those scores. In other words, Norris suggested shifting the focus from evaluating validity of a test instrument towards validity of score interpretations.

Chapelle (2012) referred to Kane (2012) when discussing challenges of providing test construct validity. She saw theoretical constructs for language tests not as ‘a priori existing entities, but rather are constructed at the interface of prior work, conceptual possibilities and pragmatic needs’ (2011: 24).

The three other types of validity – *concurrent*, *empirical* and *face* validity – seem to cause much less discussion and debate than construct validity. *Concurrent validity* is mostly seen as a degree of correlation between students’ test scores with their scores on other tests (Alderson, 1995; Richards, 1997; J.D.Brown, 2000; Hughes, 2003). Empirical, or predictive, validity is usually treated as correlation between students’ test scores with their scores on tests taken some time later (Alderson, 1995; Heaton, 1995; Hughes, 2003; H.D.Brown, 2004). Face validity seems an only validity type that ‘involves an intuitive judgement about the test content by people whose judgement is not necessarily ‘expert’:

‘The judgement is usually holistic, referring to the test as a whole, although attention may also be focused upon particular poor items, unclear instructions or unrealistic time limits’ (Alderson, 1995: 172).

Table 4.1 summarises definitions of concurrent, empirical and face validity provided by different authors in the field of language testing (Alderson, 1995; J.D.Brown, 2000; Hughes, 2003; H.D.Brown, 2004; Bachman, 2010).

Table 4.1. Concurrent, empirical and face validity as seen by different authors

<i>Validity type</i>	<i>Definition</i>
Concurrent	The degree to which a test correlates with some other test which aims to measure the same skill, or with some other comparable measure of the skill being tested. Concurrent validity is seen as correlation between <ul style="list-style-type: none"> • students’ test scores and their scores on other tests • students’ test scores and teachers’ ranking or any other such form of independent assessment • test scores and other measures (e.g. self-assessment)
Empirical (predictive)	A measure of test validity, arrived at by comparing the test with one or more criterion measures. Empirical validity can be seen through correlation between students’ test scores and <ul style="list-style-type: none"> • their scores on tests taken some time later • their success on final examination

Empirical (predictive) cnd	<ul style="list-style-type: none"> • other measures of their ability taken at the same time or some time later • success of later placement
Face	The degree to which a test appears to measure the knowledge or abilities it claims to measure, based on the subjective judgement of an observer.

Another type of validity emerges when potential consequences of a test/assessment are discussed. *Consequential* validity has been gaining attention of researchers since the early 2000s (H.D.Brown, 2004; McNamara, 2006; Davies, 2011; Kane, 2012), especially as high-stakes assessment has gained ground. According to Brown (2004), consequential validity encompasses ‘all the consequences of a test, including ... its impact on the preparation of test-takers, its effect on the learner, and the intended and unintended social consequences of a test’s interpretation and use’ (2004: 26). A similar idea had been previously expressed by Messick (1989) before consequential validity became quite a widely discussed issue of language testing:

‘if the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardised’ (1989: 88).

Kane (2012), developing ideas of importance of consequential validity, claimed that ‘researchers and test users do have an obligation to examine and report on the consequences of ... testing programmes’ (2012: 14). He set out 3 questions about social consequences of a test:

1. *What kind of social consequences should we focus on?*
2. *How should we evaluate these consequences?*
3. *Who should be responsible for evaluating consequences?* (2012: 14).

In other words, low consequential validity is treated here not as an inner characteristic of a test but rather as possible low content, construct or concurrent validity, and the effects that such low validity can have on test takers or other stakeholders.

Davies (2012) argued with Kane’s opinion and stated that validation procedures, if carried out properly, meant that likely consequences of a test had already been examined (2012: 41). Davies claimed that researchers had no control over test users, and once ‘we allow unintended consequences to be laid at the door of the researcher, responsibility loses all its meaning (2012: 41). There seems to be much reason in what Davies says about researchers having no control of test users, especially for high-

stakes international language examinations like TOEFL® or Cambridge ESOL examinations. At the same time, any test results are supposed to be used further in test takers' life, be it a small-scale diagnostic test or an exit examination at the end of the course of studies, so low validity of such tests might lead to future invalid decisions.

4.1.2. Reliability of language tests

Similarly to validity, the concept of *test reliability* and its types seem to be quite well-established issues. Reliability is seen as an essential test characteristic by many authors (Alderson, 1995; Heaton, 1995; Lumley, 2002; H.D.Brown, 2004, Ling, 2014; Yan, 2014; Kuikken&Vedder, 2014). Reliability has been discussed profoundly in publications on language testing for last 20 years.

Some authors see reliability as an independent test characteristic whereas others treat it more like a means of providing test validity. As Heaton (1995) stated, for a test to be valid, it must first be reliable as a measuring instrument (1995: 162). Alderson (1995) viewed reliability mostly as consistency of markers' work, especially those who assess writing and speaking, and their ability to give sound judgements (1995: 128).

Richards (1997) defined reliability as

'a measure of the degree to which a test gives consistent results. A test is said to be reliable if it gives the same results when it is given on different occasions or when it is used by different people' (1997: 314).

Hughes (2003) viewed test reliability in a similar vein, stating that a test is reliable if it measures consistently, i.e. one can be confident that 'someone will get more or less the same score, whether they happen to take it on one particular day or the next' (2003: 3).

He emphasised the importance of designing and administering tests in such a way that

'scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability but at a different time' (2003: 36)

H.D.Brown (2004) saw reliable tests as consistent and dependable (2004: 20), i.e. yielding similar results if administered to the same cohort of students on 2 or more different occasions. Brown saw several factors that might threaten test reliability, with the major ones being 'human error' caused by inconsistent rater work and unreliable test administration (2004: 21-22).

Reliability is widely discussed as crucial to all types and ranks of tests and examinations: from progress classroom tests to high-stakes public examinations. As a complex phenomenon, reliability may be addressed by considering a number of factors:

- reliability of test administration (Hughes, 2003; McNamara, 1997), or test/re-test reliability (Heaton, 1995: 162) that presupposes that a test yields similar results if administered on different occasions to the same cohort of test takers;
- rater reliability (Alderson, 1995; H.D.Brown, 2004; Hughes, 2003), or mark/re-mark reliability (Heaton, 1995: 162) that denotes the extent to which the same marks are awarded if the same papers are marked by two or more different examiners or the same examiner on different occasions;
- student-related reliability (H.D.Brown, 2004), which some authors view within reliability of test administration;
- reliability of the test itself (H.D.Brown, 2004): unreliability takes place if the test is a source of measurement error due to its length, poorly written items or bad timing.

Reliability of test administration mostly depends on the conditions that are created for test takers. This factor becomes especially important if the same test is administered to different cohorts of test takers at different times or in different places (e.g. international language tests like TOEFL® or IELTS that are administered in different parts of the world, or school-leaving examinations) (Heaton, 1995; Alderson, 1995, Brown, 2004, Bachman&Palmer, 2010). Heaton (1995) suggested measuring this type of reliability by re-administering the same test after a lapse of time, assuming that no test takers would receive additional training/practice between the first and second administration. Comparison of the results of the two tests, he suggested, would show how reliable the test has proven (1995: 163). The effectiveness of such a method might be seen as doubtful as there are too many factors that could interfere with test results, starting with the memory factor, when test takers might benefit from doing the same tasks for the second time, and also test takers' motivation, health, etc. Alderson (1995) considered exam administration as a key stage for providing exam reliability and, similarly to Heaton (1995: 164) stated that valid exam tasks might yield unpredictable results in case of poor administration.

Rater reliability in general and *reliability of scoring* in particular has been widely researched theoretically and empirically within the last 25 years. Alderson stressed the importance of training for all examiners, especially those who assess speaking and writing. He suggested that

'It is important that a candidate's score on the test does not depend on who marked the test, nor upon the consistency of an individual marker: an unpredictable examiner is the one who changes his/her standards during marking, who applies criteria inconsistently or who does not agree with other examiners' mark' (1995: 128).

Stressing the importance of examiners and raters, Alderson was among the first to classify reliability of test scoring into inter-rater and intra-rater. He described *inter-rater reliability* as degree of similarity of opinions of different examiners and their ability to give the same marks to the same sample of performance. Alderson admitted that 'though there is bound to be some variation between examiners and the standard some time, there must be a high degree of consistency overall' (1995: 129). *Intra-rater reliability* was seen by Alderson as consistency in the work of each examiner, i.e. the same marks being given to the same sample of performance on different occasions (1995: 128).

H.D. Brown (2004) also distinguished between inter- and intra-rater reliability and saw them as quite a common issue for classroom teachers (2004: 21). According to Hughes (2003) and Brown (2004), low rater reliability is more typical of classroom tests, rather than high-stakes examinations, although it can be observed in some standardised tests, too. H.D. Brown saw reasons for low rater reliability mostly in human factor that manifests in lack of attention to scoring criteria, general inattention, fatigue, preconceived bias or simple carelessness (2004: 21).

Although investigation of rater behaviour attracted researchers in the 1980s-90s (Charney, 1984; Huot, 1993; Hamp-Lyons, 1996), the issue of rater reliability has been quite intensively studied recently (Lumley, 2002; Huhta, Alanen, Tarnanen, 2014; Yan, 2014; Ling, Mollaun & Xi, 2014; Kuiken & Vedder, 2014). Some examination bodies also investigated the process of marking their candidates' oral and written performance (Special Test of English Proficiency (Australia), 2002; ETS (USA), 2014; Cambridge ESOL (UK)).

Lumley (2002), studying the process of rating written language performance, saw the issue as 'still not well understood' (2002: 246). Lumley aimed to investigate how raters

made their scoring decisions when marking texts written by candidates for the Special Test of English Proficiency. The examination under Lumley's study has a high stakes status because it is a part of visa application process for prospective immigrants in Australia. Lumley's research cast light on what raters actually do with the scoring categories they consider, in particular the extent to which the raters act in a similar way to each other, and if such behaviours influence rating outcomes (2002: 249). Lumley's study, based on extensive amount of empirical data from 4 raters marking 2 sets of 24 texts, revealed that although the raters understood the rating category contents similarly in general terms, they 'appear to differ in emphases they give to various components of the scale descriptors' (2002: 266). Raters' judgements appeared to be based on some complex feeling about the text, rather than the scale content, but 'they somehow managed in each case to refer to the scale content' (2002: 263). Lumley emphasised the importance of descriptors for articulating and justifying the scoring decision and, therefore, using the rating scale validly and reliably. The research also revealed another pre-requisite for successful rating - training and orientation for raters:

'Raters do not stop, as a result of training, having expert reactions, complex thoughts and conflicting feelings about texts they read (...). However, they know that they have a particular job to do and, therefore, with the benefit of training, they just cope with this demanding task, shaping their natural impression to what they are required to do, in as conscientious a manner as possible, and using the scale to frame the descriptions of their judgements' (2002: 268).

Knoch (2010), who studied behaviour of raters, found that rater behaviour does differ and depends on several factors. Such factors, according to Knoch, lead to rater variability: 'It is clear that raters attend to the rating scale criteria but not in the uniform way. (...) a great deal of rating behaviour is fixed, depending on the background of the raters or their individual rating styles (2010: 181). Knoch saw rater variability as unavoidable, like in many other situations when a human factor is involved. Knoch's study demonstrated that such differences might have a considerable impact on test results and, consequently, threaten test reliability. Similarly to Lumley (2002), Knoch saw rater training as a most efficient way of overcoming construct-irrelevant variance:

'Raters may differ in terms of their overall severity relative to other raters in the group, they may display individual biases with respect to certain aspects of the rating situation or they may vary in terms of their internal consistency or in their use of the rating scale band levels. Because rater variability is such a serious source of construct-irrelevant variance, rater training is commonly employed to limit such variation' (2010: 180).

Kuiken&Verde (2014) also investigated the process by which raters make their scoring decisions. Their aim was to look into the relationship between general measures of oral and written performance in a foreign language and overall judgements of oral and written performance by raters (2014: 280). In their research, they focused on rater behaviour, rater consistency and rater judgements, in an attempt to see what aspects raters took into account when rating linguistic performance of SL speakers and what aspects should be taken into account. The data obtained through rater interviews and self-evaluation demonstrated that, in spite of raters' effort to judge all texts according to the same criteria, they acknowledged that their expectations for lower level and higher level students were different (2014: 341). This resulted in raters attaching more importance to communicative adequacy (content, use of arguments, organization, style) than to linguistic complexity (grammar, vocabulary²², accuracy) although all criteria had equal weight. Kuiken and Verde discovered that 'raters considered the use of good arguments and general comprehensibility of a text more important, especially at lower proficiency levels'. They also observed that the raters did not focus on one specific feature at a time but combined various factors in their final judgement (2014: 342).

Ling, Mollaun and Xi (2014) investigated factors that influence raters' work, when raters determine a score for 'each constructed response' (2014: 479), both written and oral. Presuming that human raters are always trained to provide accurate, fair and reliable ratings based on scoring rubrics and guidelines, the authors stated that rater performance might be influenced by 'construct-irrelevant factors' (2014: 480) – task complexity, task type, rater background and training experiences. Apart from that, Ling, Mollaun and Xi claimed that raters feel tired towards the end of the shift, with their rating quality decreasing:

'scoring responses places a consistent burden on raters' concentration and cognitive processing ability. This can lead to time-related fatigue and threaten scoring accuracy and consistency throughout the scoring day/shift' (2014: 481)

An empirical study was conducted that involved 72 raters scoring speaking responses of the TOEFL iBT. Ling, Mollaun and Xi (2014) found that both rating productivity and quality vary greatly across hours: the 6-hour shifts had greater rating accuracy,

²² Although not stated directly, it can be presumed that 'grammar' and 'vocabulary' stand for range of grammar and lexical means, whereas 'accuracy' stands for presence/absence of errors

greater hour productivity and greater rating consistency across time than the 8-hour shifts. More than half of the raters taking part in the study felt fatigue in the afternoon, and ‘substantially more raters reported fatigue during scoring in an 8-hour shift (...). Most raters reported more re-listening behaviours and less confidence towards the end of the shift’ (2014: 494).

An insight in raters’ work that was undertaken by several researchers demonstrates that raters’ behaviour depends on many factors. These factors can be classified into 1) factors directly related to the process of rating and 2) ‘construct-irrelevant’ factors. The first group involves rating scales employed and approach to assessment (holistic::analytic), task types assessed and testees’ expected performance. The second group of factors includes raters’ previous experience, biases and attitudes; the amount of time spent marking, etc. To avoid or reduce negative effects of rater variability, all reviewed authors suggest substantial rater training and orientation before marking takes place.

As can be seen from above, rater performance can be a serious threat to test reliability, if not dealt with in a proper way. This is one of possible reasons why a number of test designers prefer closed item types that do not involve human personal judgement or assume that no test can be reliable enough. According to Hughes (2003), the best way out is to take all possible steps to provide test reliability and, at the same time, admit that some threats will still exist:

‘Human beings are not like that [they can be influenced by a lot of factors]; they simply do not behave in the same way on every occasion, even when circumstances seem identical. It implies that we can never have complete trust in any set of test scores. (...) This is inevitable and we must accept it’ (2003: 36)

Apart from rater performance, many authors, when speculating on the issue of language test reliability and ways of providing it, dwell upon other factors that threaten reliability. Heaton (1995) singled out 2 major factors that might affect the reliability of a test: the extent of the sample material selected for testing, and the administration of the test, i.e. providing the same conditions to different groups of test takers at different times. Heaton saw another dimension of test administration reliability in the quality of test materials, e.g. recordings for listening comprehension, quality of test papers, etc. (1995: 162).

In addition to the factors suggested by Heaton (1995), Hughes (2003) saw possible 'origins of unreliability' of a test in:

1. interaction between the person taking the test and the test itself, i.e. variation in the scores a person gets on a test, depending on when they happen to take it, what mood they are in, etc.;
2. scoring of the test, i.e. variation in the scores given to the same sample of performance by different markers, or to the same sample of performance by the same marker on different occasions (2003: 3-4).

Thus, threats to test reliability seem to fall into several categories, in accordance with reliability types. For this research, threats are seen as

- threats caused by inappropriate test administration;
- threats of inconsistent rater performance;
- threats caused by improper quality of test materials.

Threats to test administration include improperly described administration procedures or those procedures not observed. This might result in different test conditions for different test takers – timing, conditions in the test room, number of examiners and examiner behaviour improperly specified, etc.

Inconsistent rater performance might manifest itself in lower inter- and intra-rater reliability due to lack of rater training, lack of attention or negligence, improperly written assessment scales, lack of rater experience, etc.

Speculating on possible threats to language test reliability, Hughes (2003) suggested some practical steps that could help to avoid some of these threats. Similarly to Heaton (1995), Hughes saw those practical steps in addressing various types of reliability – reliability of administration, reliability of marking and reliability of the test itself:

Steps to provide reliability of test administration:

- provide uniform and non-distracting conditions of administration;
- identify candidates by number, not name;
- make candidates familiar with format and testing techniques

Steps to provide reliability of test:

- take enough samples of behaviour;
- write unambiguous items;

- provide clear and explicit instructions;
- ensure that tests are well laid and perfectly legible;
- do not allow candidates too much freedom, i.e. define expected outcomes;
- use items that permit scoring which is as objective as possible²³

Steps to provide reliability of scoring:

- provide a detailed scoring key;
- train scorers;
- agree acceptable responses and appropriate scores at outset of scoring;
- employ multiple, independent scoring.

(based on Hughes, 2003: 46-50)

Research conducted within last 5 years (e.g. Kuiken&Verde, 2014; Ling, Mollaun &Xi, 2014) emphasised importance of using proper rating scales and training of raters who assess both spoken and written production.

Table 4.2 presents a summary of reliability types and threats to each of them, alongside with possible ways of neutralising those threats.

Table 4.2. Major threats to language test reliability: summary

Reliability type	Major threats
<i>Reliability of test administration</i>	Administration procedures unclearly defined
	Equal conditions are not created for all test takers (timing, quality of test materials, conditions in the test room)
	Assessment criteria unclearly defined/ no assessment criteria
	Task types that are not familiar to test takers are used
	No coding of test takers' names
<i>Reliability of marking</i>	Markers/raters are not trained
	Low inter-rater reliability due to lack of rater training/experience; unclearly designed rating scales; negligence
	Low intra-rater reliability due to fatigue, lack of time, unclearly designed rating scales
	Unclearly defined expected performance of test takers
	No multiple scoring/double marking
<i>Reliability of test itself</i>	Test rubrics are not clear/allow for ambiguity
	Test materials (including audio recordings) are of poor quality
	Test is not based on a representative sample of behaviour (language means, language skills, etc.)

²³ This step can be helpful in a wide range of testing situations, where objective item types are appropriate (e.g. multiple choice tasks for testing reading, listening or vocabulary and grammar). Nevertheless, objective item types are highly unlikely for testing productive skills, where open-ended tasks often seem to be an only option

Validity and reliability are seen by many researchers as two key characteristics of language tests. As Heaton (1995) put it, test validity and reliability as ‘two chief criteria for evaluating any test’, whatever theoretical assumptions underline it (1995: 164). Davies (2011) stated that ‘reliability gives form to a test, validity gives it its meaning’ (2011: 38). Traditionally reliability has been regarded as a separate quality of a test, although nowadays some authors tend to see it as a component of validity. As Hughes (2003) wrote, ‘to be valid a test must provide consistently accurate measurements. It must therefore be reliable’ (2003: 50). Davies, developing this idea, wrote:

‘Curiously, the relationship between validity and reliability is one way: the higher the test’s reliability, the greater possibility for validity, but if one could demonstrate that a measure has good validity, its reliability can be assumed...’ (2011: 38)

However, Heaton and other authors saw the relationship between validity and reliability as a fundamental problem of testing. In contrast to the views of validity and reliability contributing to each other, an opposite situation is often observed: the greater the reliability of a test, the less validity it has (Heaton, 1995; Lumley, 2002; Hughes, 2003). A valid test may turn out to be not reliable due to various reasons, from poor administration to poor marking. A reliable test might not be valid at all, according to Hughes (2003):

‘In our efforts to make tests reliable, we must be wary of reducing their validity. ... it was admitted that restricting the scope [of a production task] might diminish the validity of the task. If we are interested in candidates’ ability to structure a composition, then it would be hard to justify providing them with a structure in order to increase reliability’ (2003: 50).

Although there are approaches to test validity that view reliability as not quite essential (e.g. Lynch, 2003), recent research demonstrates that both characteristics are still viewed as important (Lumley, 2002; Davies, 2011; Kuiken&Verde, 2014; Ling, Mollaun&Xi, 2014). It is considered essential to devise a valid test first of all and then to establish ways of increasing its reliability. This may be done by various means, with carefully designed rating systems becoming crucial. Rating systems include design of assessment/scoring scales, description of expected performance, training and support of raters.

4.1.3. Authenticity and practicality in language testing

Authenticity is a characteristic that has caused quite limited argument and has been treated as widely as ‘imitation of circumstances’ (Hughes, 2003) or as ‘the degree of correspondence of the characteristic of a given language test task to the features of the target use task’ (Bachman, 1996: 23). McNamara, when speculating on the issue of language test authenticity, claimed that a task should be close to reality. Referring to Fitzpatrick and Morrison (1971), McNamara singled out 2 dimensions of task authenticity:

- comprehensiveness, i.e. involvement of different aspects of a situation;
- fidelity, i.e. how fairly and adequately various aspects of a test task represent similar tasks in real life.

Lewkowicz (2000), after Widdowson, emphasised the importance of distinguishing between authenticity of input and output. Lewkowicz referred to the concept of test authenticity when speculating about tasks that test takers were exposed to. The major concern was those tasks not representing real life target language situations. Whereas earlier researchers such as Widdowson (1978) considered the quality of task outcome as a major dimension of task authenticity, Lewkowicz supported the idea of crucial importance of ‘authentic stimulus material’ (2000: 45), or authentic task input. Being realistic about limited possibility of providing real life tasks in language tests, the author suggested a checklist that could contribute to enhancing task authenticity:

- what degree of correspondence is needed for test tasks and target language use tasks to be perceived as authentic?
- to what extent can/do test tasks give rise to authentic-sounding output?
- does a perception of authenticity affect test-takers’ performance? (2000: 50-52).

In some way similarly to Lewkowicz, H.D.Brown (2004) suggested some practical guidelines for evaluating authenticity of language test tasks:

- the language of the test is as natural as possible;
- test items are contextualised rather than isolated;
- topics are meaningful for the learner;
- test tasks represent, or closely approximate, real-world tasks (2004: 28).

Although the usefulness of this checklist could be argued because the questions involve quite a lot of subjective perceptions (e.g. ‘language as natural as possible’; ‘meaningful topics’ very often depend on the teaching/testing situation), they added to understanding of authenticity as

- authenticity of task itself (test rubric);
- authenticity of task input, including input texts, visual sources, etc.;
- authenticity of performance (output), i.e. degree of correspondence between performance at the examination and in real life.

Test practicality/feasibility is not considered a test characteristic by all reviewed authors, with some of them treating it mostly as a dimension of context of test design but not a feature of the test itself. In this case, practicality is seen as a relatively cheap cost of test design. More researchers though (Alderson, 1995; H.D Brown, 2004, Hughes, 2003), treat test practicality wider and include, apart from effort of task design, cost of test administration (including appropriate timing and number of people involved) and transparent, specific and time-efficient scoring procedure:

‘A test which is prohibitively expensive is impractical. A test of language proficiency that takes a student 5 hours to complete is impractical – it consumes more time and money than necessary to accomplish its objective... A test that takes a few minutes for a student to take and several hours for examiner to evaluate is impractical for most classroom situations’ (2004: 19-20).

Although some difference can be observed in the ways some concepts are interpreted, for example, definitions of concurrent validity (p.54), common understanding of basic concepts has been achieved. In this research, the major parameters of test evaluation are seen as:

- test *validity*: content, construct, concurrent, and also consequential and face validity;
- test *reliability*: scoring and administration;
- test *authenticity* and *practicality*.

The issue that is considered more complicated than others in this research is evaluation of construct validity of the Final Language Examination under study. Construct validity is seen in this research, after Alderson (1995), Heaton (1995) and Brown (2004), as extent to which a test (exam) is based on its underlying theory and how

capable it is in measuring specific characteristics in accordance with that theory (whether it is a theory of language or language learning).

For quite a number of contemporary language tests, such an underlying concept is often communicative language competence (e.g. Common European Framework of Reference, 2001). For job-oriented language tests like the one under study, professional language use becomes a focus of assessment. Evaluation of construct validity of such tests would depend, in some way, on the clarity of definition of language competence for professional settings.

As demonstrated in Chapter 3, no consensus on language teacher language competence has yet been achieved. Therefore, Final Language Exam evaluation has to be based, on the one hand, on the current view of language competence and elements of teacher language competence pointed out by researchers (Chapter 3) and, on the other hand, on the data from Language Needs Analysis performed within the Exam ‘target audience’ – future and practising teachers of English who take the Exam to obtain teacher qualifications.

4.2. Direct and indirect testing

Validity, reliability, authenticity and practicality, being essential characteristics of general language tests, are also crucial for the so-called ‘performance assessment in occupational contexts’. This type of assessment is different, according to McNamara (1997), from other types of performance assessment because of ‘simultaneous role of language as a medium/vehicle of performance, and as potential target of assessment itself’ (1997: 8).

McNamara, following Jones (1985), has written extensively on performance assessment. He suggested 3 types of performance tests in occupational contexts:

- *direct assessment* that takes place directly in the workplace;
- *work sample method*, when assessment takes place in the workplace but through a set of controlled/standardised tasks;
- *simulation techniques*, when test tasks ‘involve some degree of abstraction from reality’ (1997: 43-45).

The typology above brings in the issue of *direct* and *indirect* testing. One of the first definitions of direct tests was given by Clark (1975) to refer to test formats that duplicate the setting and operation of real-life situations in which proficiency is normally demonstrated. Although the terms ‘direct’ and ‘indirect testing’ are used quite often, not much explanation can be found on the difference between them. Even in seminal publications on language testing (Alderson, 1995; Bachman, 1996; Bachman&Palmer, 2010; Hughes, 2003; McNamara, 2006), the authors drew a line between these two types of testing without further description of possible task types, assessment criteria or administration procedures. According to McNamara (1997) and H.D.Brown (2004), direct testing involves test takers in actually performing a target task, whereas in indirect testing, test takers do not perform the task itself but rather a task that is related in some way (H.D.Brown, 2004: 24).

Shohamy (1994) conducted a large-scale empirical research on direct and semi-direct oral tests – Oral Proficiency Interview (OPI) and Semi-Direct Oral Proficiency Interview (SOPI). The research focused on the task types employed in those tests and the output they elicited. Shohamy demonstrated that the direct test (OPI) stimulated ‘more natural and varied responses, i.e. fluid boundaries between the topics and smooth shifts’ (1994: 117) and also variety of linguistic and extra-linguistic means that the test takers employed. On the contrary, the semi-direct test elicited ‘sharp shifts from topic to topic and more formal language means’ and almost no extra-linguistic features (1994: 118). Shohamy did not aim to identify the best form of test, knowing that both had their advantages and restrictions, although the direct form of oral interview in her research seemed to have provided more opportunity for authentic input and output.

McNamara (1997) referred to terms ‘strong’ and ‘weak’ when discussing advantages and disadvantages of direct and indirect testing. Table 4.3 presents various aspects of strong and weak forms of tests, as described by McNamara.

McNamara saw advantages and disadvantages in both forms of performance tests. Discussing strong performance tests in general and assessment criteria that those tests employ, McNamara referred to Jones (1985). McNamara’s point was about the number of assessment criteria and their weight in strong performance tests: whether test takers’

language performance was as important as overall task achievement and whether ‘it was possible for some examinees to compensate for low proficiency by astuteness in other areas’ (1997: 41).

Table 4.3. McNamara’s view of strong and weak performance tests

	Forms of performance tests	
	<i>Strong form</i>	<i>Weak form</i>
Assessment focus	Fulfilment of the task (with language ability not always assessed)	Target language proficiency (task fulfilment is not always assessed)
Tasks	Real-world task	Specially designed tasks that may look artificial
Expected performance	Task achievement	Language performance within the scope of a task
Assessment criteria	Depend on test task type, language assessment criteria are not essential	Assessment criteria reflect language aspects (accuracy, fluency, scope, variety, etc.); ‘overall task fulfilment’ is often included, too
Types of tests based on this form	direct tests work sample	specially designed language tests (may include the occupational component)

(based on McNamara, 1997: 41-48)

Language performance on ‘weak tests’, according to McNamara, does provide a clear picture of test takers’ target language development, although

‘...one must be modest about any claims one may wish to make that tests ... provide information on the ability of candidates to communicate successfully/effectively in the workplace’ (1997: 41)

As can be seen from the above, both strong and weak forms of performance tests have their advantages and threats. Some of the reviewed authors (Alderson, 1986, Harmer, 2002, 2007) treat strong forms of performance tests (direct tests) as more valid, especially as far as content and construct validity are concerned. At the same time, such tests are more difficult to administer and mark (Slater, 1980; Shohamy, 1994; McNamara, 1997), and assessment scales are more difficult to design in comparison to ‘pure’ language assessment in weak performance tests (Slater, 1980).

Shohamy (1994) suggested criteria that could be helpful in choosing a more suitable form of a test, when there is a choice between direct and indirect testing:

- accuracy – reliability and validity of a test;
- utility – whether a test ‘serves practical information on needs of a given audience’, and also if there is possibility of proper rater and tester training;

- feasibility – whether a test is feasible to administer in a given context;
- fairness – whether a test is conducted ‘legally and ethically’ (1994: 120).

The criteria presented by Shohamy (1994) are in good keeping with the key test characteristics reviewed in part 4.1: validity, reliability, authenticity and practicality, whether the test in question is direct or indirect. For both direct and indirect forms, the issue of balancing validity and reliability is seen as one of most essential. This means first of all, choice of relevant content based on adequate sampling and job analysis (Alderson, 1995; Heaton, 1995; McNamara, 1997; Brown, 2004).

The next step is seen by many as defining test layout that includes the choice of item types used. There seems to be quite a considerable difference between item types that can be employed in direct and indirect assessment. This is the issue to be reviewed in part 4.3.

4.3. Language test formats

Review of literature on language testing demonstrates that there are certain requirements to tests that must be met – validity, reliability, authenticity and practicality (part 4.1). Nevertheless, both theoretical and empirical research show that quite often one test characteristic makes another more difficult to observe. For example, maintaining test authenticity might threaten its reliability; high reliability does not always make a test valid; high validity might make a test impractical or non-authentic. Such a situation was described by Heaton (1995) who suggested observing a balance of test characteristics and not neglecting one for the sake of another. According to Heaton, the whole language test design procedure is a set of compromises between what is ideal (desirable) and what is practical (1995: 24).

Choice of test tasks is seen by many (e.g. Shohamy, 1994; Heaton, 1995; Lewkowicz, 2000; Hughes, 2003; H.D.Brown, 2004; Bachman&Palmer, 2010) as one of the key steps in maintaining the balance between test validity, authenticity and reliability. Due to a vast amount of books, articles and research reports on testing in general and test item types in particular, only key publications in the area were reviewed in this chapter. There are several classifications of language test item types, and some similarity of

opinions is observed in how various authors treat them. Alderson (1995), Heaton (1995), Richards (1997), H.D.Brown (2004), McNamara (1997, 2007) distinguish between objective and subjective test items.

According to Heaton (1995), the difference between subjective and objective items lies in the way they are scored. In objective items ‘a testee will score the same mark no matter which examiner marks the test’ (1995: 25). Objective test items have only one or a limited number of correct answers and, therefore, can be scored mechanically. Richards (1997) considered objective items as those that ‘require the choice of a single correct answer’ (1997: 254). In quite a similar key, Hughes (2003) saw the difference between objective and subjective items in scoring procedures and, namely, in the presence or absence of human (personal) judgement in the marking process. Bachman and Palmer (2010) used the term ‘selected response tasks’ where a range of elicited responses is generally quite small and fixed (2010: 333).

The reviewed authors tend to be unanimous both in the definitions of objective items and their classification. Thus, the following item types are usually treated as objective:

- multiple choice
- error recognition/correction items;
- completion items that include gap-filling and ‘addition’ items;
- transformation items;
- matching and combination items (Alderson, 1995; Heaton, 1995; J.D.Brown, 2000; Hughes, 2003; H.D.Brown, 2004; McNamara, 2007).

Multiple choice item type is the one that traditionally has been receiving most attention and criticism. In the 960-70s it was one of the most frequently used test items, even to perform tasks for which it had never been intended, for example, for assessing writing skills. Nowadays, it is still considered by many as one of the most widely employed objective item types, although its limitations have been recognised. Frequent use of multiple choice items can be traced through key publications on language testing, examination syllabi and exam paper samples, including those for international language examinations (TOEFL®, Cambridge ESOL examinations) and some national examinations (e.g. National Examination on Foreign Languages in Russia).

The most obvious advantage of this item type, according to Hughes (2003) is that scoring can be reliable, rapid and economical (2003: 76).

The chief criticism of the multiple choice items has always been the fact that it does not allow for testing language as a means of communication. What test takers are expected to do is to choose one out of 4 or 5 options, and it bears no resemblance to using language in real life (e.g. Alderson, 1995; Heaton, 1995; H.D.Brown, 2004; Bachman&Palmer, 2010). This concern also finds proof in Hughes' opinion. Considering applicability of multiple choice and its use in language tests, Hughes emphasised a possibility of this item type giving an inaccurate picture of student performance. According to Hughes, good performance in a multiple choice grammar test is not an indicator of good performance in the productive use of grammar (2003: 76).

Despite their limited applicability, multiple choice items are considered quite useful in various testing situations that aim to assess knowledge of grammar and vocabulary, i.e. 'to recognise correct forms and to make important discriminations in the target language' (Heaton, 1995: 27).

Good quality multiple choice items are considered very difficult to construct. The key sentence (stem) is expected to be brief, on the one hand, and, on the other hand, contain enough information for test takers to make right choices. There should be only one correct option. Although this might seem an easy task, many authors emphasise great difficulty of constructing multiple choice items (e.g. Alderson, 1995; Heaton, 1995; Hughes, 2003; Bachman&Palmer, 2010). Items must have only one correct option (key), with other options looking attractive and not providing any possibility of pure guessing for candidates.

Error recognition items often have the form of multiple choice, where test takers need to identify one incorrect form out of 4/5 underlined. This task type was very widely used in the paper-and-pencil and computer versions of TOEFL²⁴ in 1990-2000s (<https://www.ets.org/toefl/pbt/about>, retrieved on February 7, 2015). The major criticism of this item type, in addition to the general criticism of multiple choice items, is that recognition of error does not always mean ability to produce correct items (e.g.

²⁴ This task type is no longer employed in the modern version of TOEFL – TOEFL iBT

Alderson, 1995; Heaton, 1995; J.D.Brown, 2000; Hughes, 2003). This criticism is supported by some psychologists (e.g. Anastasi, 1988; Балыхина, 2006; Nihae, Chiramanee, 2014) who claim that it is undesirable for learners to be exposed to wrong items, even if the task is to identify the incorrect element. However, according to Heaton, this item type is closely related to the skills that students need for writing, checking and proof-reading their essays, reports or articles (1995: 40). For teachers, an ability to spot mistakes in written and oral performance gets paramount importance.

Matching and combination items usually aim to test the ability to select appropriate responses. This may include finding proper answers to the given questions, matching a term and its definition, finding pairs of synonyms/antonyms, matching words and transcriptions, etc. Matching items test mostly recognition, be it grammatical, phonological or lexical units. The expected response of test takers is minimal, and this is one of the major reasons for the limited applicability of this item type. Nevertheless, this item type is employed relatively frequently, with the Teacher Knowledge Test (Cambridge ESOL) providing a good example of matching tasks (Spratt, Pulverness, Williams, 2006). In each module of the examination, up to 25% of tasks are based on the ‘matching’ principle, although the design of those tasks may be slightly different (<http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/tkt/>, retrieved on February 8, 2015).

Completion test items, known also as gap-filling items, are seen as those that measure ‘production rather than recognition’ (Heaton, 1995: 43) and test the ability to use (produce) the right language form to fill in the gap, but not just choose the form from the given options. Due to this reason, in some tests completion items are preferable to multiple choice items.

Completion items are easier to design, compared to multiple choice items, but more difficult to mark and always require human markers (Heaton, 1995). Sometimes even the most straightforward completion items can cause problems in scoring due to different interpretations of the item by designers and test takers. Therefore, design of completion items should always take into consideration ‘the length to which the test writer must sometimes go to make certain that testees produce only the answer he or she wants to be used in each blank’ (Heaton, 1995: 43).

Transformation items are seen as useful for testing production – mostly grammar and vocabulary. Heaton (1995) sees this item type as the one that ‘helps to provide a balance when included in tests containing multiple-choice items’ (1995: 46). The reason for such an opinion can be found in the nature of transformation: being an objective item type, it comes closest to items that test production - like essay or paragraph writing (part 4.3.2 of the current chapter).

Similarly to completion items, transformation items can elicit more than one correct answer. Therefore, being not very difficult to design, they cause extra effort for considering all possible variants of answers and including those variants in the key to the task. Even with carefully designed keys, transformation items make mechanical scoring quite challenging, and often involve human raters.

Table 4.4 presents the major objective task types that are employed in contemporary language testing, with the advantages and disadvantages of each type summarised.

Table 4.4. Objective task types in language testing

Test item type	Advantages	Disadvantages	Applicability
<i>Multiple choice</i>	Easy and quick to mark; marking can be done mechanically	Difficult to design Requires minimal (non-verbal) response from test takers Does not assess ability <i>to use</i> the target language Chance of guessing the correct answer	- grammar (recognition) - vocabulary (recognition) - reading comprehension - listening comprehension
<i>Error recognition</i>	Easy and quick to mark; marking can be done mechanically	Tests recognition, but not an ability to use the target language Exposes test takers to incorrect language form	- grammar (mostly recognition) - possible for testing vocabulary
<i>Matching</i>	Quite easy to design Quick and easy to mark; may be marked mechanically	Requires minimal (non-verbal) response from test takers Tests recognition only	- grammar (recognition) - vocabulary (recognition) - listening - reading
<i>Completion items (gap-filling)</i>	Test production, though limited in scope Relatively easy to design, although much effort is invested in avoiding misinterpretation and providing all possible options in answer keys	May be quite difficult to mark; always require human raters Sometimes allow for different interpretations and more than one correct answer	- grammar (production) - vocabulary (production) - reading comprehension - listening comprehension

Test item type	Advantages	Disadvantages	Applicability
<i>Transformation</i>	Tests production, though expected performance is limited in scope Not very difficult to design	Takes a lot of effort to design keys with all possible responses Takes quite a long time to mark Always requires human raters	- grammar (production) - vocabulary (production)

(based on: Alderson, 1995; Heaton, 1995; H.D.Brown, 2004; Hughes, 2003, McNamara, 2006)

Objective item types are quite often seen as those that are simpler to answer than subjective ones (Alderson, 1995; Heaton, 1995; Hughes, 2003; McNamara, 2006) and even allow for some chance of wild guessing, especially the types based on recognition rather than production (multiple choice, matching). Another criticism is that objective items have rather limited applicability and cannot be used for testing productive skills.

As Heaton (1995) put it,

'It should never be claimed that objective tests can do those tasks which they are not intended to. (...) they can never test the ability to communicate in the target language, nor can they evaluate actual performance' (1995: 27).

Despite some criticism, objective test items are seen as more 'universal' in terms of their applicability, and they can be used for testing various skills, knowledge and language elements. Nowadays they are mostly employed for testing receptive skills – reading and listening, as well as grammar and vocabulary, although in the 1960s there were attempts to use objective items for assessment of productive skills, especially writing. Few modern tests consist entirely of multiple choice or other objective items. A big number of tests, including international language examinations, 'strike a happy balance' (Heaton, 1995: 33) between objective and subjective items, so that command of grammar and vocabulary as well as the ability to use the target language productively are in the focus of assessment.

Speaking and writing skills are usually tested by *subjective items* that require test takers to perform writing and speaking tasks similar to those required in real life. In contrast to objective items, subjective ones are not always easily classified, and often authors prefer to discuss them while discussing particular issues of writing and speaking assessment. Parts 4.3.1 and 4.3.2 give an insight into various ways of

employing objective and subjective item types in assessment of communicative skills
– listening, reading, speaking and writing

4.3.1. Assessment of receptive skills

Assessment of listening and reading is an issue that has been getting quite considerable attention from various researchers – from those investigating language testing issues and designing language tests (Alderson, 1995; Bachman, 1996; Bachman&Palmer, 2010; Fulcher&Davidson, 2006; Douglas, 2014) to the authors dealing with wider issues of language teaching (Ur, 1996; Nunan, 2002; Harmer, 2007; Thornbury, 2006). In this research, only seminal publications are reviewed in an attempt to design a taxonomy of assessment task types that can be employed for assessment of language teacher language competence. The taxonomy based on literature review is supplemented by a taxonomy of tasks used by national and international examination bodies for assessment of teacher language knowledge and skills (part 4.5.2). These taxonomies have a dual aim in this research. First, they are used for the evaluation of the format and administration of the current Final Language Exam. Second, the taxonomies are employed in designing alternative examination tasks (Chapter 13) and discussing various ways of changing the current Final assessment practices.

According to listening types, tasks for testing listening are often classified into:

1. tasks for testing *selective listening* (H.D.Brown, 2004), or *listening for specific information* (Harmer, 2007; Ur, 1996), when test takers process discourse such as a short monologue or a conversation (news, weather forecast, directions/instructions) in order to get information on names, figures, places, or certain facts and events. These tasks aim to assess the ability to ‘scan’ for required information, rather than ability to fully understand and interpret a stretch of discourse;
2. tasks for testing *intensive listening*, or *listening for detailed understanding* that aim to assess the ability to process longer stretches of discourse for full understanding;

3. tasks for testing *extensive listening* that aim at assessment of the ability to process longer samples of spoken language (lectures, TV programmes, conversations, etc.)²⁵.

The 1st group of tasks usually includes

- listening and choosing (providing) appropriate response – gap-filling, choosing the right picture/map/description, etc. These tasks are sometimes called information transfer tasks, with ‘classical’ examples being ‘Listen and fill in the chart’ or ‘Listen and choose the right picture’, etc. Some authors call these tasks ‘aural scanning’ (H.D.Brown, 2004: 129), when a listener selects relevant pieces of information from the text they hear.
- ‘listening cloze’ (H.D.Brown, 2004: 125), or partial dictation, when test takers listen to a story, a monologue, or a conversation and simultaneously read a written text in which some words/phrases have been deleted. This type of listening task is considered to have a number of weaknesses. Weaknesses embrace *lower validity* of such tasks, when both listening and reading are tested in fact, and success of listening depends not only on listening, as the task declares, but also on reading skills. Another weakness of this task type – *lower reliability* - can be observed when test takers are supposed to fill in the gaps in the reading text with exactly the same words as in the listening text. This makes scoring easier, because it can be done mechanically using a key, and the task may be considered close-ended. However, this is not always possible in reality (Alderson, 1995; Heaton, 1995; Hughes, 2003) because test takers may fill in the gaps with words/phrases with the same meaning as those used by the speaker, but not exactly with the words they hear in the text.

The 2nd group of tasks usually includes ‘responsive listening tasks’, when test takers are expected to choose (multiple choice or matching) or provide responses (gap-filling or answering open-ended questions) to the given statements/questions, or mark items as true or false. Sometimes tasks of this group include ‘information transfer’ tasks,

²⁵ Within extensive listening, some authors (e.g. Heaton, 1995; H.D.Brown, 2004) single out listening for gist, or the main idea, without fully understanding every detail. Other authors (e.g. Ur, Harmer) see listening for gist as a separate type of listening, although they see the nature of listening for gist in a similar way

when test takers are expected to provide non-verbal responses to the text: draw a picture/map, identify the person in the photo, put the pictures in the right order, etc.

The 3rd group of tasks is quite similar to the 2nd one but is performed with longer texts. Some authors (e.g. Heaton, 1995; H.D.Brown, 2004) single out another subgroups of tasks here, which they call authentic tasks: note-taking, editing and retelling (oral or written). These tasks are extensively used in proficiency tests (TOEFL and IELTS) to assess candidates' ability for academic listening.

Tasks with open-ended responses, from short answers to open-ended questions, to note-taking and editing, may seem easier to design than multiple choice or gap-filling tasks that need to meet a lot of requirements. Nevertheless, there are some threats that open-ended tasks have to face. The first one is the degree of 'freedom' (Hughes, 2003) that test takers are given. If test takers are expected to produce written or oral response, the task must say clearly how extended this response must be, if test takers are expected to use the same words as used in the listening text. The latter might become a threat to the task validity. The reason for this is what Heaton called 'a memory factor', when a listening test becomes more of a memory test:

'We are rarely called upon to remember the exact words someone spoke in real life unless in very unusual circumstances, e.g. evidence given in a court case, in which a speaker's exact words may have great significance. Even in such circumstances, individuals have great difficulty in recalling the actual words spoken even though they can remember perfectly the general meaning of what the person said. Therefore ... avoid setting questions which involve the memorisation of individual words in sentences' (1995: 86).

The second threat of open-ended listening tasks is that scoring of open-ended responses might become problematic and question both validity and reliability of the task. Thus, rules must be set whether spelling and grammar mistakes are taken into consideration, whether range of language means (grammar, vocabulary) is important in open-ended responses. If so, according to Heaton (1995), Hughes (2003), H.D.Brown (2004), a task becomes an integrated task of listening and writing/speaking with a set of assessment criteria including not only degree of text understanding, but also content and accuracy of open-ended response.

The study of key publications on *testing reading* revealed some similarity of opinions concerning reading types and tasks for assessing them (Alderson, 2000; Hughes, 2003; H.D.Brown, 2004; Thornbury, 2006; Harmer, 2007). As said in the introduction to this chapter (p.51), the selection of sources is far from being exhaustive with only key publications on teaching and testing reading reviewed. The choice of sources for review in this research was informed primarily by their focus – assessment of reading in general and teacher reading in particular, although the latter area has not attracted much attention of researchers.

Reading is often seen as selective reading for specific information (scanning), intensive and extensive reading, and also reading for gist (skimming). Some authors (H.D.Brown, 2004; Richards, 1997; 2010) see reading aloud as one of reading types, and such an approach is implemented in some of language examinations for teachers (e.g. LPATE examination in Hong Kong, reviewed later in this chapter). Other authors (Heaton, 1995; Ur, 1997; Nunan, 2002; Hughes, 2003; Harmer, 2007) do not see reading aloud as a separate type of reading and, therefore, do not include it in the range of reading types subject to assessment.

Tasks for assessing reading are seen as very much similar to task for testing listening and classified into:

- *close-ended tasks*: multiple choice, true/false statements, matching, gap-filling with verbal and non-verbal (picture-cued) responses;
- *cloze tasks* that are often seen as a separate type of tasks although its form resembles gap-filling;
- *open-ended tasks* that vary from short-answer questions to note-taking, text editing and information transfer, with the latter involving both verbal and non-verbal input.

Another group of tasks for reading assessment can be singled out – *integrated tasks* that comprise two or more skills, for example, reading and speaking or reading and writing (Heaton, 1995; Alderson, 2000). Although it can be argued that such tasks tend to have higher authenticity in comparison to the task types listed above, their use in some tests (e.g. TOEFL) is quite limited. This might be explained by validity threats caused by involving several skills (part 4.1), when performance on a reading task turns out to depend on skills other than reading.

Similarly to tasks for testing listening, tasks for testing reading can be based on various types of input – verbal and non-verbal, with the latter comprising pictures/ series of pictures, diagrams, graphs and tables, maps, etc. According to Heaton (1995), the sampling of input for reading tasks is of the utmost importance and must be related to the broader aims of the language teaching situation, i.e. to fit the aims of target language teaching:

'Ideally, in a test of proficiency the test should contain the type of reading task which will be demanded of the testees in later real-life situations' (1995: 118)

The range of texts that can be employed for testing various types of listening and reading seems to be vast – from advertisements and timetables to pieces of fiction (Heaton, 1995; Alderson, 2000; H.D.Brown, 2004; Harmer, 2007). The range of texts for assessing listening and reading skills of teachers of English as a FL is discussed further in this chapter (part 4.5), when national and international language examinations for language teachers are reviewed.

4.3.2. Assessment of productive skills

From the point of view of assessment, the major issue of speaking is its close interrelation with listening that often makes it difficult to ‘isolate oral production tasks that do not directly involve the interaction of aural comprehension’ (H.D.Brown, 2004: 140). According to Brown (2004), only in limited contexts – speeches, retelling, storytelling or reading aloud²⁶ can oral ability be assessed without the aural comprehension involved. If test takers’ speaking performance is affected by effectiveness of their listening skills, validity and reliability of oral tests can be compromised and even threatened. Thus, one of the biggest challenges of oral task design is ‘teasing apart, as far as possible’ the factors/variables related to poor listening comprehension and those related to poor speaking skills.

Another challenge seen by many (Heaton, 1995; Bachman, 1996; Hughes and Palmer, 2003; Bachman&Palmer, 2010) is the design of oral assessment tasks, or elicitation

²⁶ As far as reading aloud is concerned, it can be argued that, although it does not involve listening comprehension, it involves reading comprehension: successful reading aloud can be highly problematic without sufficiently developed reading ability

techniques which are all open-ended, although expected output can differ in scope. These techniques, according to Brown (2004) should be designed so that they, on the one hand, provide test takers with choice but, on the other hand, elicit exactly those target forms (grammatical, lexical, phonological) that are expected from testees.

Closely connected with open-endedness of oral task is the issue of scoring. Whereas for receptive skills the expected performance can be predetermined and limited by the task, tasks for assessment of productive skills can only predetermine expected *range* of performance, but not the exact output. This makes scoring more challenging, compared to objective assessment tasks, and always involves human raters and subjective judgements (part 4.1).

Many of these issues have to and can be addressed through adequate choice of oral production tasks. Several taxonomies have been suggested by different authors. Heaton (1995) suggested a classification based on type of input (verbal and visual) and output (limited vs extended). H.D.Brown (2004) employed a classification of speaking sub-skills as a basis for oral task taxonomy. For this research, oral tasks are classified and reviewed in accordance with types of speaking. Types of input that is employed in various task types were also classified (Figure 4.2, p.83).

In the table below, oral tasks are presented as those assessing monologue and dialogue. An attempt has also been made to arrange the tasks in the order of increasing scope (the expected output) and degree of spontaneity.

<i>Monologue</i>	Picture description (based on visual input only, or both visual and verbal) Comparing and contrasting two or more pictures Giving directions/instructions based on visual input The short talk (Heaton, 1995: 102) Picture-cued storytelling Presentation Retelling a story/an event
<i>Dialogue/ conversation</i>	Dialogue completion and oral questionnaires ²⁷ (Brown, 2004: 149-150) Interview Discussion and conversation/conversational exchanges Role play Games (Brown, 2004: 175-176)

²⁷ Test takers are presented a dialogue with one interlocutor's lines being omitted. They have time to read it through and then respond to the statements, with the role of the interlocutor being played by an examiner or being reproduced as a recording

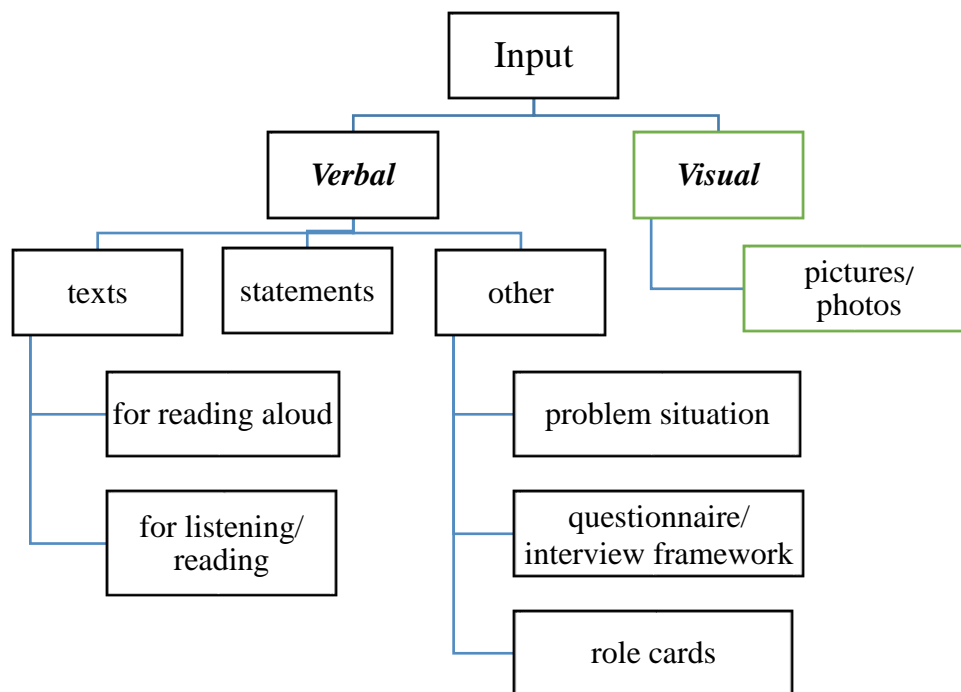
Tasks for assessment of oral performance can be divided into 2 groups – those for assessing monologue and dialogue (conversation) skills. Some authors (Heaton, 1995; H.D.Brown, 2004; Richards, 2010; Coniam&Falvey, 2013) consider reading aloud as a separate type of speaking. No agreement has yet been reached on whether reading aloud should be considered a speaking task. Some authors see it as a variation of prepared monologue (Heaton, 1995; H.D.Brown, 2004) whilst others (Alderson, 1995; Bachman, 1996; Richards, 1997; Bachman&Palmer, 2010) consider reading aloud to be outwith the range of speaking skills due to a different nature of this task.

Much has been written about advantages and disadvantages of various task types. The reviewed authors examine validity and reliability of the task types above and look at different ways of dealing with threats. Threats can be caused by tasks themselves, for example, by the degree of freedom given to test takers (Hughes, 2003) and, therefore, test takers' performance meeting expected performance requirements. Some task types involve more than one skill (e.g. reading aloud) but are still treated as tasks for speaking assessment, which, in some cases might compromise their validity. Discussions and conversations that involve more than 1 person might put reliability at risk because they involve a lot of factors other than participants' speaking ability, with many of them being psychological and social ones. The status of role plays as assessment tasks is questionable: some authors (e.g. Heaton, 1995; Brown, 2004) consider role plays suitable for oral assessment, others²⁸ see this task type as only suitable for continuous and peer assessment but not for formal tests. The major reason for not including role plays in the formats of formal and high-stakes examinations seems to be involvement of factors other than speaking abilities, for example, interaction styles, temperament, ability to act spontaneously.

²⁸ There is a vast amount of publications on role plays and their use in development of speaking skills in TESOL which are not reviewed in this research. Most authors see roles as beneficial for creating positive atmosphere in the classroom and reducing inhibition, especially with beginner learners. Role plays are also seen as means of providing real communication in the language classroom. In other words, role plays are seen as a good opportunity to practice meaningful language but not to assess it.

All of the task types above are based on some sort of input, which can be verbal, visual or both. Figure 4.2 summarises various types of input that can be employed in oral assessment tasks.

Figure 4.2: Types of input for oral production tasks



Selection of input, according to Heaton (1995), Hughes (2003), Brown (2004), Bachman and Palmer (2010) must follow certain principles. Thus, a framework for an effective interview contains a number of mandatory stages: warm up, level check, probe and wind-down (Canale, 1984). Choice of visual stimuli should be based on expected performance and elicit exactly what is in the focus of assessment (Hughes, 2003).

In contrast to objective task types with one or a limited number of correct answers, assessment of speaking always involves human judgement. As discussed earlier in this chapter (part 4.1.2.), this requires human raters, a set of assessment criteria and rating scales (Heaton, 1995; Lumley, 2002; H.D.Brown, 2004; Kuiken& Vedder, 2014). The review of literature revealed a common approach to assessment criteria, although different authors refer to different assessment scales and emphasise importance of different criteria. The major set of criteria can be seen as:

- accuracy (phonological, grammatical, lexical)

- fluency
- task achievement (accomplishment)
- range (phonological, grammatical, lexical).

Common European Framework of Reference (2001), in addition to the criteria above, suggests ‘content’ and ‘scope’, i.e. how extensive test takers’ performance (output) is. For integrated tasks, ‘comprehension’ (H.D.Brown, 2004) is added. As claimed earlier in this chapter, integration of skills within one task may influence the task validity. Therefore, despite the fact that integrated tasks tend to bear more resemblance to real life, some authors (e.g. Hughes, 2003) feel like avoiding them in assessment.

Whatever assessment criteria are employed, they involve design of rating scales, a thorough description of expected performance, training and monitoring of raters. As presented earlier in part 4.1.2, rater performance is considered a key issue in providing reliability of any test, and especially those that involve extensive responses being marked. However, only few authors (e.g. Fulcher&Davidson, 2006) emphasised the importance of *interlocutor training*, which can be seen as essential step in providing exam reliability, alongside with rater consistency. Improper interlocutor training may result in their spontaneous behaviour (e.g. offering clues, asking additional questions, developing discussion if the topic seems interesting, etc.) and, therefore, unequal conditions to test takers.

Assessment of writing ability includes assessment of handwriting, spelling, and writing as a productive skill. According to Brown (2004), assessment of *handwriting* is based on a limited range of techniques that test the ability to produce letters and symbols: copying, writing numbers and abbreviations, converting numbers and abbreviations to words, form completion (2004: 222-223). Although Brown was writing in 2004 and treated handwriting as a skill of paramount importance, not much changed within last 10 years and not many authors concentrate on it when discussing issues of assessment.

Spelling tests has been seen by many as an integral part of writing assessment (e.g. Heaton, 1995; Hughes, 2003; Brown, 2004). In contrast to assessment of productive writing (see further), tasks for testing spelling can be classified as objective and very often close-ended. This often allows for mechanical scoring and does not involve

human rater judgement. Spelling tests may take a form of a ‘traditional, old-fashioned dictation’ (Brown, 2004: 223), matching, and multiple choice techniques.

Writing as a productive skill may be subdivided into controlled (display) and or ‘real’ writing (Brown, 2004: 225), or, as other authors put it, writing as a means and writing as an aim (e.g. Harmer, 2007; Thornbury, 1997; 2006). Writing as a means may aim to test spelling, and also grammar and vocabulary when test takers are offered transformation tasks, tasks that involve production of sentences and short texts based on verbal and visual clues.

Assessment of ‘*real*’, or productive, writing is based on a range of tasks – from close-ended tasks like paraphrasing or writing short answers to questions, to open-ended tasks that expect extensive performance – letters, reports, summaries, essays. In this part, only this type of writing and issues of its assessment are reviewed.

Assessment of writing has occupied attention of various authors and resulted in an extensive range of publications: books, articles, dissertations. Alderson (1995), Heaton (1995), Hughes (2003); H.D.Brown (2004), McNamara (2006) dealt with fundamentals of testing writing, from types of writing activities to scoring written performance. More recent research seems to have accepted the existing typology of writing tasks and deals more with issues of administration and scoring of written responses, training of raters and other ways of providing reliable results, especially in high-stakes testing (Lumley, 2002; Barkaoui, 2010, 2013; Kuiken&Verde, 2014; Huhta, Alanen, Tarnanen, 2014).

Agreement has been observed in how various authors treat tasks for assessment of productive writing. The taxonomy below was designed on the basis of several key publications on issues of teaching and assessment of reading skills (Nunan, 2002; Harmer, 2004, 2007; Alderson, 1995; Hughes, 2003; H.D.Brown, 2004). Thus, tasks are classified in accordance with the expected product (a report, summary, etc.) and the input (verbal or visual):

- *postcards, cards, forms*

- *letters and emails*²⁹, both informal and formal
- *reports* (book reports, project summaries, laboratory reports) that test ability to convey the purpose/main idea; logical organisation of the text; ability to conclude; ability to choose appropriate vocabulary;
- *summaries* (lectures, videos, etc.) that test the ability to understand the main idea; omit unnecessary detail; convey thoughts logically; use quotations when required;
- *narration, description, argument* that assess the ability to state the purpose; use effective language means (both accuracy and range) for purpose achievement;
- *interpretation of statistical or other graphical data* that tests attention to detail, ability to compare and contrast data/sources and accurately present the results in writing;
- *essay and composition*

Similarly to assessment of speaking, assessment of productive writing is based on a quite a complex system of scoring with human raters involved. H.D.Brown (2004), Bachman&Palmer (2010), Fulcher, Davidson&Kemp (2011) consider 3 major approaches to scoring: holistic, primary trait and analytical. According to Brown (2004), in the first method, a score is assigned to a sample of performance (a letter, summary, essay, etc.) which represent the marker's general overall assessment. This system of scoring is considered by many as the one quite easy to use (Barkaoui, 2010). However, some authors tend to be cautious about applying holistic scoring to speaking and writing tasks. Fulcher&Davidson (2006) stated that

'in this kind of evaluation the argument that the score adequately summarises the evidence depends upon the acceptance of a collective understanding of the meaning of the descriptors. There must be a group of people whose ability to place language samples into categories has evolved over time, and into which newcomers can be socialised' (2006: 96).

Even those who prefer holistic scoring for its ease of use have to admit that the system has disadvantages because it aims 'to achieve high inter-rater reliability at the expense of validity' (Weigle, 2002: 114).

Primary trait scoring can be seen as a variation of holistic scoring, with only one factor being rated (e.g. accuracy only, or logical structuring only, etc.). This type of scoring

²⁹ Email writing is becoming a part of some English examinations, e.g. in Finland, when test takers write 2 emails – a personal one and a formal one to the manager of an online retailer (Huhta, Alanen, Tarnanen, 2014)

might seem not reliable enough because, concentrating on one essential criterion (e.g. achievement of the communicative purpose), it, in some way, neglects others, like accuracy or layout. However, some authors see such scales as useful because they determine priorities for assessment in situations when it is impossible or undesirable to take into consideration all assessment criteria.

Analytical scoring is based on a set of criteria, with a score being given for each (e.g. a score for linguistic accuracy, a score for layout, etc.). Bachman&Palmer (2010) also see analytical scoring systems as sets of criteria, with the latter defined by the specific constructs to be measured (2010: 341).

Each scoring method has its advantages and disadvantages that are summarised in Table 4.5.

Table 4.5: Advantages and disadvantages of different scoring methods

Task type	Advantages	Disadvantages
<i>Holistic/global</i> (Bachman &Palmer, 2010)	Fast evaluation Relatively high inter-rater reliability Applicable to all writing tasks (Brown, 2004)	Raters need to be thoroughly trained to perform scoring One score masks differences across the subskills within each score (Brown, 2004) Raters may weigh the hidden components differently in arriving at their single rating (Bachman, 2010: 341) May threaten reliability (Fulcher and Davidson, 2006)
<i>Primary trait</i>	Allows both writer and evaluator to focus on one primary feature	By concentrating on one criterion, neglects others Features other than the primary one can be implicitly evaluated and influence the score
<i>Analytical</i>	More detailed than holistic scale Tend to reflect what raters actually do when rating samples of language use (Bachman&Palmer, 2010: 342) Rating of each component is more explicit, weighting is easier to control Allows for washback by calling attention to the areas that need improvement Suitable for both progress and final assessment Can be easily changed or adapted	One scale may be not suitable for all types of writing It may be difficult to ‘assign levels’ in scale design (Bachman, 2010: 340)

(based on: Alderson, 1995; Hughes, 2003; H.D.Brown, 2004; Bachman & Palmer, 2010; Fulcher&Davidson, 2006;)

Whatever scale is employed for assessment, decisions on final scores are always made by human raters (see part 4.1.2). This is one of the reasons why rating scales are sometimes seen as inefficient and unreliable ways of scoring language tests, with preferences given to objective (selected response) items. Bachman&Palmer (2010) admitted that concerns about potential rater reliability and subjective human scoring were very serious, but claimed that they were ‘by no means insurmountable’ (2010: 352). Bachman and Palmer suggested the following steps to be taken for the rating procedures becoming highly consistent and relatively efficient:

- *anticipating problems* that can be caused by high demand on human resources, i.e. involving more raters to provide multiple scoring, to enlarge the total amount of rating time, especially for rating longer samples of performance:

‘This demand on human resources must be recognised as an unavoidable cost of obtaining the kinds of information that ratings can provide [task effectiveness, task impact on test takers]. ... we believe that the potential gains in meaningfulness and generalizability more than offset any potential loss in practicality’ (2010: 352).

- *dealing with inconsistency*, which can be attributed to 3 causes: different interpretation of scales, different standards of severity, and reaction to elements not relevant to scales. Bachman&Palmer (2010) suggest several ways of minimising the negative effect of rater inconsistency:
 - preparing raters (see also part 4.1.2) through appropriate training (discussion of scales, reviewing language samples that have already been rated; rating language samples and discussing the scores given; monitoring time each rater spends for marking)
 - obtaining a sufficient number of ratings, i.e. providing multiple rating
 - estimating the reliability of the ratings while designing test tasks

(Bachman&Palmer, 2010: 351-353)

4.4. Language testing for language teachers: theoretical and practical considerations

In contrast to issues of general language testing that have been receiving a considerable amount of attention since the 1960s, language testing of teachers of English as a FL has gained much less interest. There are several publications that present theoretical and empirical research in the area of design and implementation of language examinations for language teachers in different countries. Grant (1997), Elder (2000) and Consolo (2008) dealt with issues of developing examinations for teachers of FL (Spanish, Italian, Japanese) in English speaking countries – the USA and Australia, and the examination for teachers of English as a FL in Brazil. The authors see such examinations as very important for teacher development: within the existing range of language examinations assessing proficiency in FL, there is no ‘examination of language communicative proficiency developed and widely used specifically for teachers’ (Consolo, 2008: 1).

Grant (1997) and Elder (2000), with a three years’ difference, emphasised that the model of language teacher language competence was far from being thoroughly described. As Elder wrote, ‘proficiency prerequisites for language teachers are all too often defined quantitatively, in terms of ‘seat time’ or hours of formal study’ (Elder, 2000: 1).

Whilst Grant suggested using the model of Communicative Language Ability (Bachman, 1990), Elder suggested specific language skills required for teaching purposes:

- command of subject-specific/metalinguistic terminology;
- discourse competence required for classroom delivery of subject content, i.e. command of linguistic features (directive, questioning, rhetorical signaling devices, simplification strategies, etc.) (2001: 154).

The examinations for language teachers described by these authors (reviewed further in part 4.4) – the Spanish test for bilingual teacher certification in Arizona (Grant, 1997) and Language Proficiency Test for Teachers (LPTT) of Italian and Japanese (Elder, 2001) – aim at assessing FL teacher language communicative skills in listening, reading, writing and speaking with a special emphasis laid on Classroom Language

(Elder, 2001), Interaction with parents (Grant, 1997) and teacher ability to use the target language as a medium of instruction including the ability ‘to produce well-formed input’ and ‘to draw learners’ attention to the formal features of the target language’ (Elder, 2001).

Similarly to Alderson, H.D. Brown, Hughes and McNamara, Elder and Grant view test validity, reliability and authenticity as key characteristics of performance tests for teachers. Grant (1997) addressed issues of validity and reliability through a series of questions about the test for teachers being discussed. Judging by these questions, the author was mostly concerned with content and concurrent validity of the test together with various aspects of reliability of results. Indirectly rather than directly, Grant addressed the issue of construct validity. As discussed previously in part 4.1.1, construct validity is seen as a complex dimension of language testing. In testing for FL teachers, construct validity seems to be even more difficult to argue due to absence of a commonly accepted definition of language teacher language competence. Thus, for defining construct validity of a language test for FL teachers Grant (1997) suggested asking for opinion of different ‘contributors’ – teacher trainers, practicing and expert teachers and researchers (1997: 38) – on what constitutes language competence required by teachers.

The issue of validity raised the issue of authenticity which can be treated in different ways but mostly was viewed as authenticity of test tasks and input (part 4.1.3 of this chapter). Referring to Bachman (1991), Elder distinguished between

- situational authenticity, i.e. the level of correspondence between the test and target language use situation;
- interactional authenticity – the capacity of the test task to engage the relevant language abilities of the test taker (2001: 155).

For assessment of teachers, for whom the most authentic task is teaching in the language classroom and the most authentic input is that obtained from the learners, ‘practical considerations such as time and resources do not make such a naturalistic setting an option’ (Grant, 1997: 26), and the tests in question are less situationally authentic because they are not administered in the language classroom. The language tests reviewed by Elder presupposed reading a story aloud ‘as if to a group of young

school age learners' and issuing a set of 'classroom-like' instructions for a particular learning activity (2001: 156). In other words, classroom-related situations were simulated in the classroom. Elder saw obvious limitations of such tests where 'tasks are delivered as monologue by candidates' (2001: 156).

Elder (2001) and Grant (1997) saw language testing for language teachers as a specific area of language assessment having some features of ESP assessment, with the key issues of test design remaining similar to 'general' assessment, i.e.:

- providing content and construct validity of test materials through careful selection of content and background theory of competence;
- providing authenticity of test to the highest possible degree keeping in mind test practicality, which is quite difficult to balance in language tests for teachers when authenticity requires them to be administered in a language classroom whilst practicality puts them back to formal exam surrounding;
- ensuring reliability of results through secure administration, scoring and marking procedures.

ESP features of teacher language assessment, according to Grant (1997) and Elder (2001) manifest through the selection of classroom-related content and situations; task types that would test not only an ability to use, but also to teach the target language.

One of the most recent publications that describes Brazilian experience in designing an examination for teachers of English as a FL (Consolo et al, 2008) concentrated mostly on importance of such an examination for teacher certification. The authors pointed out some studies of 'reality of EFL teachers in Brazil', which indicated lack of teacher language proficiency in teachers on the one hand and, on the other hand, lack of a commonly accepted definition of this proficiency (2008: 2). Referring to literature, Consolo suggested taking into consideration the following dimensions of teacher language:

- grammar and syntactic structure;
 - vocabulary and pronunciation;
 - fluency;
 - strategies of verbal interaction;
 - reading and writing abilities
- (Consolo, 2008: 5)

Consolo did not specify why those dimensions were chosen and what exactly each of them meant. For example, the author did not explain if ‘grammar’, ‘vocabulary’ and ‘pronunciation’ meant accuracy of use, grammatical and lexical range or both; what ‘strategies of verbal interaction’ included or why reading and writing abilities were in the list with speaking and listening missing, although later in the article (2008: 10), speaking and listening were described as test papers.

Although vague about some issues, Consolo’s publication is considered important for this research for 2 reasons. First, he raised some issues of designing language examinations for language teachers – from why such exams are essential to what they should focus on. Second, he provided an outline of such an exam by defining its format – writing, speaking and listening papers.

Apart from ‘ESP features’ (Grant, 1997) of language assessment for foreign language teachers, some current assessment practices, including the Final Language Exam under study, bring out issues of CLIL – Content and Language Integrated Learning (e.g. Short, 1993; Coyle, Hood, Marsh, 2010; Gablasova, 2014). CLIL is seen as an approach where curricular content of subjects (in the case under study, linguistic subjects like Theoretical Grammar, History of English, Lexicology) is taught to students through a language that is neither their first language nor the dominant medium of instruction in the education system (Gablasova, 2014). Apart from teaching issues like content selection and choice of appropriate teaching techniques, CLIL adds some issues to assessment, where the main problem is related to students’ mastery of the target/assessment language and the extent to which their command of the language places constraints on their ability to express the content knowledge they have.

Short (1993) emphasised the importance of distinguishing between the language and the content knowledge of students, and taking effort so that one does not interfere with the demonstration of the other. Coyle, Hood and Marsh (2010) considered assessment in CLIL ‘a major area of teacher uncertainty’ (2010: 114) and suggested several questions to be answered before choices of assessment content and format are made:

- what is assessed: content or language?
- can students perform assessment tasks in their L1?

- what tools can be used for assessment?
- provided we assess in English, how can we minimise the effect of the language in the content assessment?
- how can we evaluate subject skills and language skills? (2010: 115)

Gablasova (2014), referring to Hofmanova, Novotna and Pipalova (2008) suggests that teachers ‘may not be sure whether a student is simply unable to demonstrate knowledge because of a language barrier or whether, indeed, the student does not know the content material being assessed’ (2014: 151).

Coyle, Hood and Marsh (2010) suggested using separate sets of assessment criteria for knowledge and language skills under assessment. For assessment of knowledge acquired through CLIL teaching, they single out several levels at which learners can operate and which, therefore, can be assessed:

- factual recall
- general understanding
- ability to manipulate the content, using higher-level thinking skills such as interpretation, analysis, synthesis or application
- ability to research more independently and extend the topic knowledge beyond what has been presented by the teacher (2010: 116).

Gablasova’s research (2014) looked at the other component – language skills to be assessed, and the choice of language for assessment. Gablasova analysed advantages and disadvantages of using L1 and L2 as languages of assessment of content knowledge. Similarly to Hincks (2010) Gablasova reported more fluent and elaborate speech of exam takers when they were given an opportunity to perform assessment tasks in their mother tongue. The same learners, however, experienced difficulties with technical vocabulary in L1 if they had learnt this vocabulary in L2.

Although assessment in the target language may cause difficulties for learners to express the required content, Gablasova considers this an issue that can be dealt with by employing transparent assessment criteria, or ‘linguistic features that directly affect the quality and comprehensibility of the content knowledge (2014: 153):

- accuracy
- fluency

- appropriate academic format
- appropriate vocabulary.

Whilst teaching and assessment foci of many programmes are defined by educational institutions, a clear distinction should be made between assessment of knowledge and assessment of language. It is fully understood in this research that a division between linguistic and teacher's knowledge, and language skills employed to express this knowledge is quite vague, with a lot of interrelated areas. However, distinction between linguistic knowledge and language skills seems essential. The necessity of such a distinction manifests in the current Exam format where Task 1 and Task 2-3 have clearly different assessment foci. However, the difference in assessment foci does not lead to a difference in assessment criteria (see further, Chapter 5) with the same set being applied to all 3 tasks. This research makes an attempt to draw a line between assessment of linguistic knowledge and language skills for the case under study (see Chapter 13).

4.5. Language examinations for English language teachers: national and international experience

The study of existing national and international language examinations for language teachers of foreign languages (both pre-service and in-service levels) demonstrated that a limited number of countries have officially established standardized examinations for teachers, like Praxis® in the USA. In most cases, administration of such exams is a task performed by colleges/universities through a system of final examinations. The reviewed examinations were divided into 2 groups: language examinations for teachers administered by international exam bodies – ETS (USA) and Cambridge ESOL (UK); and national language examinations developed by particular countries – Australia, Brazil, Hong Kong, USA.

1. International language examinations for language teachers

- Praxis® - an ETS examination for language teachers used for licensing and certification processes across the USA;

- Cambridge ESOL international examinations for language teachers:
 - ICELT – In-service Certificate in English Language Teaching;
 - TKT® - Teacher Knowledge Test;

2. *National language examinations for language teachers*

- Exame de Proficiência para Professores de Língua Estrangeira - EPPL (Brazil)
- Language Proficiency Assessment for Teachers of English – LPATE (Hong Kong)
- Language Proficiency Test for Teachers of Italian [as a foreign language] (Australia)
- Arizona’s Spanish Proficiency Test (USA)

The review aimed to summarise experience in the design of language examinations for language teachers in different countries, i.e.:

- to compile a list of teachers’ knowledge and communicative skills that are assessed;
- to define content areas under assessment
- to make a taxonomy of task types employed in the examinations under study.

The aims of the review defined its structure. First, assessment foci were analysed for different examinations, with language areas and communication skills under assessment singled out. Then emphasis was laid on task types employed for assessing those skills and assessment criteria, as well as marking schemes used by different examination bodies.

4.5.1. *Assessment focus of national and international language examinations for language teachers*

The examinations under study differ in administration mode - from pencil-and-paper to online, in exam length and duration, and pre-requisite level of English expected from exam-takers from B1 (Threshold) to C1 (Effectiveness).

According to the exam syllabus, **Praxis® tests** measure subject-specific content knowledge, as well as general and subject-specific teaching skills of a language teacher. Part 1: ‘*Subject assessment*’ measures general and subject-specific teaching skills and knowledge. Part 2: ‘*Principles of learning and teaching*’ and Part 3: ‘*Teaching foundations*’ measure general pedagogical knowledge and pedagogy in

‘English Language Arts’ (<https://www.ets.org/praxis/prepare/materials/5195>, retrieved on 15 February, 2015). For this research, only Part 1 is reviewed as relevant.

Set of tasks 1 tests candidates’ ability to spot the mistakes in students’ oral and written texts. For this candidates listen to recordings of student talk and identify the mistake then read samples of students’ work and, again, identify the mistakes. The main focus of these tasks is language awareness, teacher listening and reading (short texts).

Set of tasks 2 checks candidates’ knowledge of phonology (transcription, articulation, intonation and stress), morphology and syntax, psycholinguistics (SLA, code-switching, student motivation, etc.) and sociolinguistics (dialects, appropriate language use, communicative competence) through a series of multiple choice tasks. The major emphasis of this section is on language teacher subject-specific knowledge with reading skills being both focus and a means of assessment.

Sets of tasks 3-5 concentrate on approaches and methods in ELT, classroom management, lesson planning; assessment and curriculum development, focusing on subject-specific and pedagogical knowledge and reading (descriptions of classroom situations or samples of teaching materials).

According to the **ICELT examination** syllabus³⁰, the examination consists of 3 components: *Component One*: Language for teachers, where candidates are required to complete four language tasks. *Component Two*: Supervised and assessed teaching, aims to assess lesson planning skills, classroom teaching skills and lesson evaluation. *Component Three*: Methodology assignments aims to evaluate candidates’ general language awareness and awareness of ELT theory together with academic skills.

The ICELT syllabus (2005) gives a detailed description of requirements to an English teacher’s communicative competence. In *speaking* and *writing* candidates are expected to:

- speak and write in language that provides a natural model for learners and that does not cause an audience to question the teacher’s professional language competence;
- speak with pronunciation which is internationally intelligible;

³⁰ <http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/icelt/>, retrieved on August 19, 2014

- write with a level of accuracy in spelling and punctuation which reflects the standard required for an English language teacher at this level;
- write well-organized texts;
- use a range of language to express themselves in a variety of styles in both social and professional contexts (2005: 15).

Professional competence in *reading* and *listening* presupposes a language teacher comprehending:

- a range of professional written material (schoolbooks, books/articles for teachers, etc.);
- a range of professional listening material, including talks on professional topics (2005: 15).

In the area of Classroom Language (spoken and written language in the classroom), a teacher is expected to:

- give oral and written instructions;
- maintain control and discipline; acknowledge and praise;
- elicit and prompt responses; ask questions;
- present and model new language; write on the board, produce accurate handouts, OHTs, posters, etc.;
- tell stories, read aloud;
- adapt/create texts, tasks and tests;
- evaluate and comment on learners' written work; write reports on learners' progress

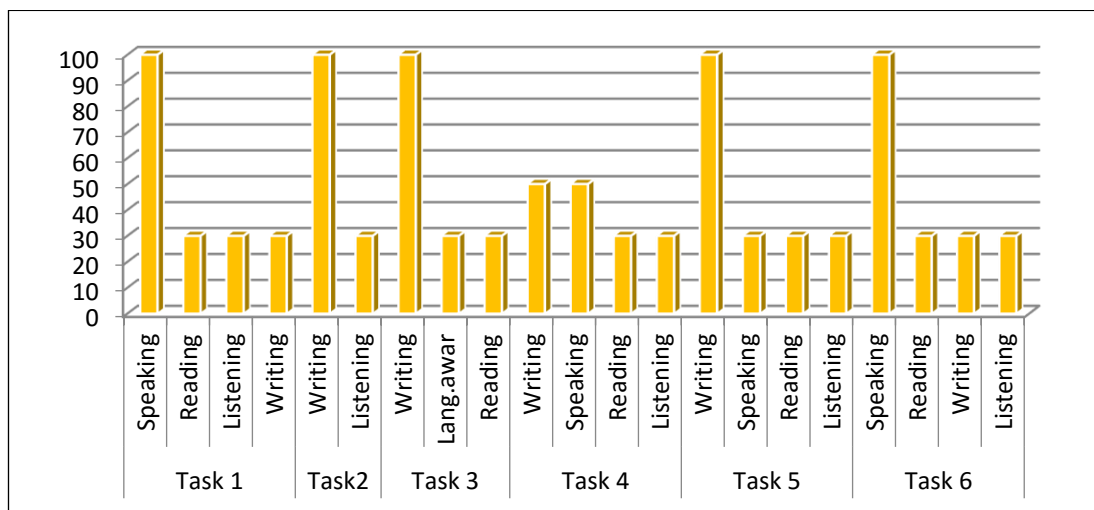
(ICELT syllabus, 2005: 9-15)

The examination is designed so that communicative skills of listening, reading, speaking and classroom English, and writing are the major focus of assessment. This can be traced through task types (reviewed further in part 4.5.2) and assessment criteria (2005: 39-40). The assessment focus of ICELT can be defined as all 4 communicative skills, classroom English, teacher language awareness and pedagogical knowledge required by a teacher of English as a FL. It should be noted that communicative skills under assessment at the examination can be viewed as:

- skills which are in the focus of assessment;
- support/pre-requisite skills that are involved in successful task completion.

Figure 4.3 illustrates the distribution of skills within exam Tasks 1-6 in ICELT, with the tallest bar (100%) for each task being the major assessment focus and other bars standing for skills involved but not assessed directly.

Figure 4.3: Communicative skills under assessment in ICELT



(based on ICELT Syllabus, 2005: 21-27)

The TKT examination (<http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/tkt/>) is a test of professional knowledge for English language teachers. This knowledge includes concepts related to language, language use and the background to and practice of language teaching and learning.

Module 1 tests candidates' knowledge of terms and concepts of ELT:

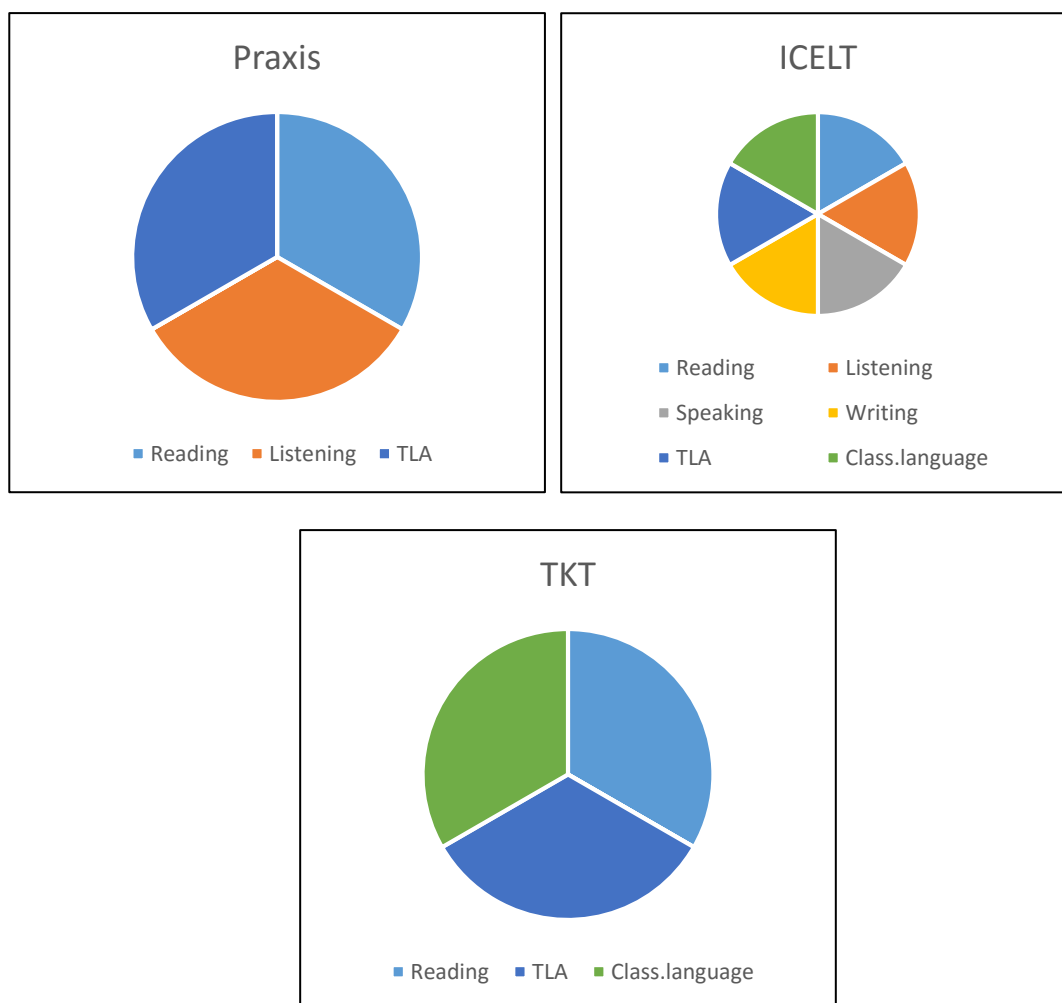
- concepts and terminology for describing language; language skills and sub-skills;
- factors that influence language learning process (motivation, differences between L1 and L2, learner styles, etc.);
- range of methods, tasks and activities for teaching and assessment.

In addition to assessment of receptive grasp of terminology, the module aims at testing candidates' reading skills. There are no tasks assessing candidates' productive use of terminology, or productive skills of speaking and/or writing.

Module 2 focuses on what teachers consider and do while planning a lesson and series of lessons. As well as in Module 1, the major focus is on professional terminology and reading. Module 3 focuses on classroom management and classroom interaction and aims at assessing candidates' ability to use English appropriately in the classroom – to give instructions, explain new language items, categorize learners' mistakes, etc.

The range of skills under assessment varies from purely receptive listening and reading of short texts (Praxis®) to writing an argumentative essay or keeping a diary on conducted classes (ICELT). Some examinations (Praxis®, TKT, ICELT) pay special attention to the command of ELT terminology, Classroom English (LPATE, ICELT) and Teacher Language Awareness (ICELT, LPATE). Figures 4.4, 4.5 summarise representation of skills and knowledge areas in various examinations under review.

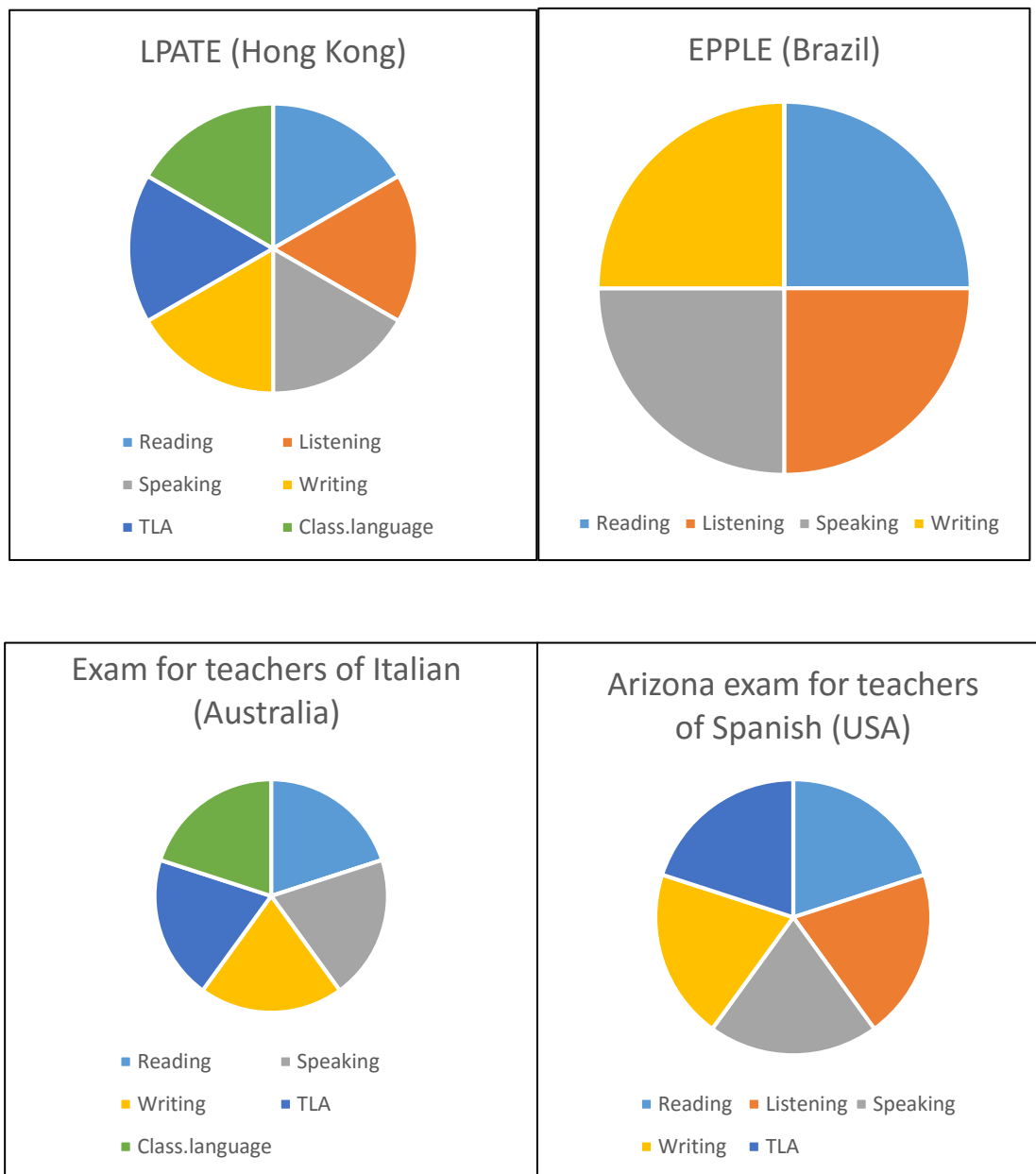
Figure 4.4: Representation of communicative skills and knowledge in international language examinations for language teachers³¹



³¹ TLA stands for teacher language awareness (Chapter 3)

The segments of the pie charts above and below do not present the balance of skills, i.e. how thoroughly each skill is assessed at each examination, but just the structure of examinations. The presence of a segment in a pie chart means that the skill it presents is assessed; if there is no segment, it means the skill is not in the assessment focus.

Figure 4.5: Representation of communicative skills in national language examinations for language teachers



As the diagrammes above demonstrate, among international examinations, ICELT is the one that embraces most skills and practices, including Classroom Language and Teacher Language Awareness (TLA). Within the range of national examinations, LPATE includes the whole range of communicative skills, whereas other examinations have slightly narrower foci. In terms of language teacher language competence and the ways it is seen by researchers (Chapter 3), the reviewed examinations deal with the majority of its components:

- receptive and productive communicative skills – listening, reading, speaking and writing that language teachers apply in and out of the classroom;
- classroom language that comprises giving instructions, maintaining discipline, explaining language items, error correction, etc.
- teacher language awareness;
- knowledge of how languages are learnt and taught.

The division of communicative skills into ‘traditional’ listening, reading, speaking and writing may be considered slightly artificial because a lot of situations involve more than one skill. However, such division follows the logic and layout of Exam materials that often consist of Listening, Reading, Speaking and Writing papers (e.g. TKT, Praxis®, LPATE, EPPLE). For example, listening tasks may resemble ‘general’ listening tasks when exam takers listen to a lecture and choose the right option for each question. However, in other task types listening can be interrelated with Teacher Language Awareness, when exam takers listen to a text and define its difficulty for a certain group of learners.

The reviewed examinations differ not only in their focus but also in the level of task difficulty – from tasks involving knowledge and recognition (e.g. Praxis®, TKT) to complex integrated tasks involving multiple skills and professional practices (e.g. ICELT, LPATE).

4.5.2. Task types employed in national and international language examinations for language teachers

Apart from a considerable difference in assessment focus (part 4.5.1) – from receptive skills of listening and reading (Praxis, TKT) to a wide range of receptive and productive skills, including classroom language and teacher language awareness (e.g. ICELT, LPATE), the examinations differ in assessment tasks. They all can be classified into *objective* (close-ended), i.e. those that require only one or limited number of right answers and *subjective* (open-ended) that presuppose a response that cannot be treated as right or wrong but should be evaluated using a set of criteria. The most widely employed close-ended tasks are:

- *multiple choice*, that presuppose choosing either one or several answers from a number of options. This task type in the reviewed examinations is widely employed for assessment of listening (Praxis®) and reading skills (Praxis®, TKT); command of vocabulary (ICELT) and teacher language awareness (Praxis®);
- *matching*, most often used for assessment of ELT terminology and knowledge of techniques and methods of teaching languages (Praxis®, TKT);
- *grouping*, employed by TKT in assessment of teacher professional vocabulary.

The range of open-ended tasks is wider in the reviewed examinations. These tasks require responses from short (one sentence) to extended (keeping a diary or reflecting on a language lesson) and can be classified into tasks assessing oral performance and tasks assessing writing skills of language teachers.

Oral performance tasks vary from interviews, when exam takers answer examiners' questions, to problem solving. A group of tasks that can be singled out within oral performance assignments is Classroom Language tasks. This group of tasks includes both *short responses* like formulating a question or setting up an activity and *extended responses* like conducting a class or evaluating teaching materials. Written tasks vary from note-taking while reading an ELT article to designing and evaluating teaching materials and keeping a diary of conducted classes. Table 4.6 summarises the task types employed for assessment of skills and knowledge in language examinations for language teachers.

Table 4.6: Assessment foci and assessment tasks in language examinations for language teachers³²

Knowledge areas and skills under assessment	Assessment task types (tasks marked in <i>italics</i> are used for continuous assessment only)
<i>Listening</i>	Objective task types: - multiple choice - matching
	Subjective task types: - watching a video fragment and its further discussion (with examiners/ in pairs with other candidates)
<i>Reading</i>	Objective task types: - multiple choice (one or several correct options) - matching
	Subjective task types: - reading and summarising a text on ELT issues - materials evaluation (e.g. evaluating difficulty of a text) - reading and identifying genre of a text - <i>conducting a small-scale research on a chosen ELT issue</i>
<i>Speaking</i>	Subjective task types: - interview (answering examiners' questions) - oral presentation of a researched issue (based on reading texts) - lesson evaluation (giving feedback on observed lessons) - problem solving (in pairs with other candidates)
<i>Classroom Language</i>	Objective task types: - multiple choice (e.g. defining an aim of instruction for an activity) - matching (e.g. instruction to its aim) - grouping (e.g. instructions according to lesson stages)
	Subjective task types: - conducting a lesson, including further reflection on Classroom language employed - materials design - giving instructions, eliciting and explaining language items, checking comprehension, dealing with misunderstanding - reciting a poem - storytelling - reading aloud
<i>Writing</i>	Subjective task types: - note taking (while reading an article/ observing a lesson) - providing feedback on student written performance - <i>materials design</i> and evaluation - syllabus design - writing a lesson plan - <i>reflecting on a conducted lesson</i>
<i>Teacher Language Awareness</i>	Objective task types: - multiple choice (recognising errors in student oral and written performance) - identifying errors in student performance - providing terms for given definitions
	Subjective task types: - giving definitions to the terms provided - materials design

³² Based on review of exam syllabi of national and international language examinations for language teachers. Continuous assessment tasks are presented in *italics*

It should be noted that assessment tasks differ significantly in the difficulty level, scope of expected performance - from choosing a right answer from a given list to producing oral/written texts, and skills involved in performing these tasks – from one-skill tasks to integrated tasks comprising reading/listening, speaking and writing. As can be seen from above, quite a substantial number of tasks are integrated, i.e. involve more than one skill and/or knowledge areas.

Key information on the reviewed language examinations for language teachers that is considered relevant for this research is presented in Table 4.7.

Table 4.7: International and national language examinations for language teachers: summary

	Potential exam takers	Administration mode	Knowledge and communicative skills tested	Expected level of 'general' English/other target language
<i>International examinations</i>				
Praxis®	in-service	paper test/ computer based test; a series of close-ended tasks (m.choice)	reading, listening, linguistic terminology, ELT terminology at the receptive level	not stated directly
TKT	pre-service, in-service	a paper test; a series of close-ended tasks	reading, teacher language awareness, ELT terminology	min B1 ³³ (Threshold level)
ICELT	in-service	a paper test + a series of open-ended tasks	speaking, listening, knowledge of terminology, reading and writing, teacher language awareness	min B2 (Vantage level)
<i>National examinations</i>				
LPATE	in-service	a paper test + a series of oral and written assignments	reading, speaking (including reading aloud), writing, classroom language; teacher language awareness	not stated directly
EPPL	in-service teachers of English as a FL in Brazil	a paper test with a computer version being developed; a series of open-ended tasks	reading, writing, listening, speaking	from 'minimally necessary to excellent proficiency in English' (Consolo, 2008)

³³ According to the system of levels presented in Common European Framework of Reference (2001)

	Potential exam takers	Administration mode	Knowledge and communicative skills tested	Expected level of 'general' English/other target language
Test for Italian teachers ³⁴	in-service teachers of Italian as a FL in Australia	oral test (30 min) – a series of open-ended tasks	reading, speaking (including reciting poems and reading aloud), writing, classroom language; teacher language awareness	sufficient for teaching Italian in primary schools in Australia
Test for Spanish teachers ³⁵	mostly in-service teachers of Spanish in the USA	oral and written test in a language laboratory – a series of close-ended and open-ended tasks	speaking, listening, knowledge of terminology, reading and writing, teacher language awareness	not stated directly

The publications in the field of language testing for language teachers (Consolo, 2008; Elder, 2000, 2001; Grant, 1997) demonstrated the authors' interest and concerns of the issues that are typical of general language testing (Alderson, 1995; Heaton, 1995; McNamara, 1997; J.D.Brown, 2000; H.D.Brown, 2004; Bachman and Palmer, 2010):

- validity of the test, including content and construct validity;
- authenticity of exam tasks;
- reliability of exam administration procedures;
- practicality of examination.

Out of these 4 issues authenticity very often comes to the fore, according to McNamara (1997) and Elder (2001). It might be explained by peculiarities of teacher job and tasks teachers perform. These tasks are quite difficult to simulate in the examination classroom. This may be one of the reasons for shifting the focus from formal examinations in the examination centre (Praxis®, TKT, EPPLE), to continuous or portfolio assessment of teachers performing in a real classroom (DELTA, LPATE, some ICALT modules). This approach brings in another issue: continuous assessment is likely to be possible for practising teachers at in-service level. Examinations designed for pre-service teachers (graduates of colleges and universities) stick to either

³⁴ Language proficiency test for teachers of Italian

³⁵ The Arizona classroom teacher Spanish proficiency exam

close-ended tasks that assess listening, reading and ELT terminology (Praxis®, TKT) or simulations in the examination room (EPPLÉ).

Closely connected with authenticity, is the issue of validity that in some way makes language assessment for language teachers problematic. Grant (1997) and Elder (2001) were mostly concerned about absence of a commonly accepted structure and description of language teacher language competence and, therefore, absence of common grounds in evaluating construct validity of language tests for language teachers. Douglas (2000), speculating on validity of professionally-oriented tests as opposed to general language tests, came to the conclusion that:

'It is proven to be very difficult, and may eventually prove to be impossible, to make predictions about non-test performance in the real target situation on the basis of a single test performance, no matter how true to real-life the test tasks may be. This is so because language use, even in highly restricted domains ... is so complex and unpredictable that coverage, or sampling of tasks, will be inadequate' (2000: 12)

Both Elder (2001) and Douglas (2000) saw a possible solution in extensive analysis of context and providing a taxonomy of skills that language teachers require. Such skills proved to be different for different countries and educational environments. This might explain the fact that the examinations described above differ significantly in their content, format and administration.

The review of 1) elements of language competence that teachers of English are expected to demonstrate (Chapter 3) and 2) test evaluation parameters (validity, reliability, authenticity, practicality) and requirements to language test design informed the design of the Final Language Exam evaluation checklist. The Checklist is based on Alderson's (1995) exam evaluation framework with some elements specified according to the context of research:

- *examination content and format* (content areas under assessment, skills and sub-skills, topics, tasks and their applicability to FL teaching);
- *exam administration* (procedure, length, timing for each section, people involved);
- *marking procedures* (keys for objective marking, rating criteria for subjective marking, grading and setting pass marks);
- how the *results* are reported (what is reported, who the reports are available to);

- *Exam design* that includes the following parameters:
 - *test specifications* and how detailed they are about test purpose, expected performance, length, format (task types, papers, rubrics), content (skills and communication areas/topics), level of difficulty, mode of delivery (paper-based, computer-based, iBT);
 - *procedures for examination materials development* (people involved, major steps taken, measures for providing validity and reliability, item moderation and pre-testing);
 - *training of the Examination staff* (examiner training, training of administrators and raters).

The parameters in the Checklist were divided into two groups: those that could be described on the basis of written/published documents and those that required stakeholders' responses. The white areas in the checklist deal with the issues which can be evaluated on the basis of the existing documents provided by the university (Final Examination Syllabus; samples of Exam materials; Dean's orders on members of the Examination Board) or obtained from the Ministry of Education web-site (e.g. State Educational Standards for FL teacher development). The grey areas required information and opinions to be obtained empirically (see Chapter 6: Research Methodology).

Table 4.8. Final Language Exam evaluation checklist

Test specifications
Is there a description of the content of the examination?
If Yes, does this include
test purpose
the description of candidates
test format (papers, rubrics, task types, length)
content (knowledge, skills, topics, sub-skills)
level of difficulty/level of expected performance
mode of delivery (PBT, CBT, iBT)
assessment criteria and marking procedure
samples of student performance on tasks
Item writing and moderation
Is there a team of writers appointed officially?
Does the team consist of university teachers only?
Can school teachers be involved in item writing/moderation/piloting?
Are item/test writers given any guidance (exam syllabus, sample tasks, papers, etc.)?
Do item writers meet at any point to discuss each paper/test?
Are the materials produced by the writers discussed within the team <i>before</i> they become exam materials?
Are the items/papers moderated?

Are the items/papers pre-tested?
Are assessment criteria (speaking, writing) piloted?
Examiner training
Do examiners and administrators have any training sessions before the examination?
Is there a standardisation meeting before the exam?
Are there any feedback sessions after the exam?
Exam content and format
Is the number of tasks the same for each candidate?
Are the following areas of knowledge tested?
linguistic theory
teacher language awareness (see Chapter 3)
grammar
professional vocabulary
general vocabulary
pronunciation
Are the following skills tested?
speaking - monologue
speaking - dialogue
listening
reading
writing
Classroom English
What topic areas are involved?
general (within Common European Framework requirements)
professional (teacher-specific)
Are the examination tasks/questions of the same level of difficulty?
Is the examination content the same each year?
Can changes be introduced easily and quickly?
Exam administration
Are administrative details clearly established before the exam?
Are exam procedures the same for each candidate?
Does each candidate get the same amount of time for preparation/answering?
Can students complete the exam reasonably well within the set time frame?
Is there any distribution of roles in the exam committee?
Marking procedures and reporting results
Is the scoring/evaluation system feasible in the given time frame?
Are assessment criteria available before the examination
to examiners?
to students?
Are there any criteria which have more weight than others?
Do all members of the exam committee take part in marking?
Are methods of reporting results determined in advance?
Are the results reported on the day of the exam?
Is there any report on the exam results?
If Yes, does the report include
the results of the exam – grades obtained
the analysis of difficulties candidates faced
statistics (e.g. task difficulty index, mean score, etc.)
comparison of candidates' results with the results demonstrated earlier
the description of skills assessed
Is the report available to all staff?
Is the report available to students?

The Checklist presents areas of data collection that are viewed as essential in evaluating Exam validity and, therefore, Exam strengths and weaknesses. Each Checklist area is seen as a means of obtaining data on Exam validity, reliability, authenticity and practicality, as shown in Table 4.9.

Table 4.9: Design of Exam evaluation checklist and data collection

Exam evaluation parameters	Information to be obtained on:
Validity	Exam specifications Exam item design Exam content and format
Reliability	Examiner training Exam administration Marking and assessment
Authenticity	Exam format Exam content
Practicality	Exam format Exam administration Exam materials design

The Checklist served a springboard for designing research instruments and obtaining data from various stakeholders (Chapter 6) that complemented the information from ministerial and institutional documents on Final Language Exam design and administration (see further, Chapter 5).

Chapter 5

Description of the current Final Language Examination for future teachers of English as a Foreign Language in Russia

This part presents a description of the current Final Language Examination (hereafter – the Examination) for the graduates of FL Department at Tula State Pedagogical University. This description is based on analyses of documents – State Educational Standards (developed at the federal level by the Ministry of Education), the Final Exam Syllabus (developed by the University, i.e. at the local level), Exam materials samples developed by the Department of Foreign Languages and documents issued by the Department (e.g. Dean’s orders on Examination board members).

Chapter 5 starts with a brief overview of the documents under analysis: their status, aims, and freedoms/restrictions they give to Exam designers, examiners and administrators. Then the Final Language Exam is described alongside the parameters presented in the Exam evaluation checklist (Chapter 4, p.107-109): its content, format and administration are given as close a look as the data permits. The description results in defining gaps in information that need to be filled through obtaining qualitative and quantitative data for answering the research questions.

5.1. Overview of documents involved in Exam design and administration

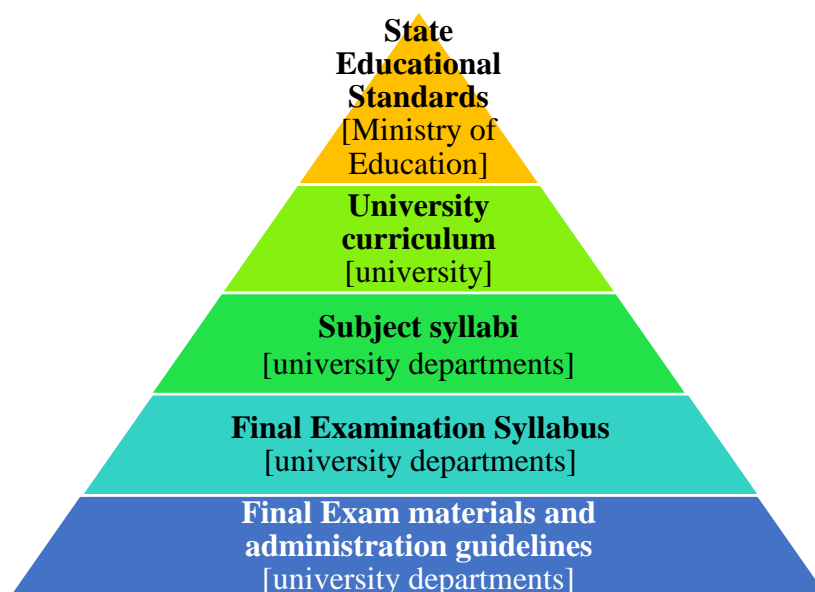
The following documents were analysed:

- State Educational Standards for Foreign Language Teacher Training to obtain information on:
 - language skills a university graduate (future teacher of English) is expected to demonstrate at the end of the course of studies;
 - a range of tasks and topics a university graduate is supposed to deal with;
 - any State requirements to the Final examination format, content, administration (if any) universities must follow;
- Requirements to university accreditation (issued by Rosobrnadzor, an affiliated structure of the Ministry of Education) to see if there are some centralized guidelines on Final exam administration, its content or format;

- the Final Exam Syllabus for university graduates (prospective teachers of English) issued by the university:
 - to obtain information on language skills and other assessment areas the Examination focuses on;
 - to compare the focus of assessment as presented in the syllabus with the State requirements presented in the State Educational Standards;
- Exam materials samples – to add to the information obtained from other sources on skills under assessment, topical areas, text and task types employed for the Final assessment.

The format and content of the Final Language Examination are defined by universities so, theoretically, different universities can have different final assessment formats. There is one requirement to be met: the content of the examination must match the content of disciplines/subjects taught throughout the course of studies, with the latter matching the content prescribed by the State Educational Standards. The hierarchy of documents involved in Final Examination design is presented in Picture 5.1.

Picture 5.1: Regulatory documents for final examinations at Russian universities



The Standards are developed by the Ministry of Education (hereafter – the Ministry, <http://eng.mon.gov.ru>) and renewed on a routine basis every five years. As a result, the universities must revise their assessment formats in accordance with new

Standards also every five years, or more often if they consider it necessary, although changes made are sometimes simply formal: changing the document number, upgrading the list of recommended sources and web-links for exam preparation, changing the name or status of the institution, etc.

The current version of the State Educational Standards for FL teacher training in Russia (<http://www.fgosvpo.ru/uploadfiles/fgos/3/20111115120152.pdf>), referred to as ‘generation three’ Standards, was issued by the Ministry in 2010. The difference between the 2010 version and the previous ones is the fact that it is the first Standard for the three-level professional education (Bachelor – Master – Doctor) which Russia adopted having joined the Bologna process in 2003. The structure and content of the new Standards differ from the previous one (version 2005) due to significant changes in the Ministry’s approach to its general supervision and quality control at universities. Thus, according to the Bologna declaration, universities are given more freedom in defining the subjects to teach and the content of the subjects together with the number of academic hours. Since 2010, universities are supposed to develop a list of competences that graduates must demonstrate at the end of the course of studies, instead of those competences being ‘prescribed’ by the Ministry in the earlier versions of the Standards. Thus, a description of competences is now seen as a part of curriculum development at university. The curriculum for FL teacher training designed by Tula State Pedagogical University (http://tsput.ru/about_us/overview/the_basic_educational_program/GOS/undergraduate.php, retrieved on June 16, 2014) aims at the following competences for future FL teachers³⁶:

- a graduate demonstrates linguistic knowledge – knowledge of phonological, grammatical, lexical phenomena and of how language works;
- a graduate is aware of accepted behaviour in target countries and models of intercultural interaction;
- a graduate can achieve various communicative aims by employing appropriate linguistic means;
- a graduate can express thoughts freely and spontaneously, employing various linguistic means;

³⁶ Translated from Russian; only language competences are presented here

- a graduate is aware of formal, neutral and informal registers of communication; can fight stereotypes and communicate in general and professional areas (FL teacher training curriculum, 2010: 5-6)

The Curriculum serves a basis for syllabus development for each subject, and designing the Final Exam Syllabus that contains:

- a description of the format of the Final Examination on Theory of Education, Psychology and TESOL methods; and Final Examinations on the 1st and 2nd Foreign Languages;
- the topics to be covered at all three Examination tasks;
- sample questions/tasks and a list of recommended sources for each examination;
- assessment criteria.

The Final Examination Syllabus for any subject, including English/German/French must be approved by the Faculty Council (an elected body of lecturers, administrators and students) and serves as a basis for examination materials design by departments – English, German and French.

The Final Examination syllabus approved by the Faculty of Foreign Languages in 2011 states that the graduates ‘take the final examination which aims at testing the graduates’ ability to teach languages to different learner groups’ (Final Examination Syllabus, 2010: 1). The assessment area is seen as broadly as

‘graduates’ language competence’, i.e.:

- *professional language skills at professional level;*
- *socio-cultural competence;*
- *an ability to use the target language for professional development*

(Final Examination Syllabus, 2010: 4)

As can be seen from the quotation above, graduates’ language competence is seen by the Syllabus designers quite broadly and vaguely. First, some confusion can be observed between ‘professional language skills’ and ‘an ability to use the target language for professional development’. It can only be presumed that the former implies knowledge and skills necessary for teaching language (e.g. linguistic knowledge, teacher language awareness, classroom language) whilst the latter means self-study skills, autonomy and self-evaluation. Second, ‘professional level of skills’ needs to be defined more precisely, using either a commonly accepted system of levels (e.g. CEFR, 2001) or providing a detailed description of knowledge, skills and activities.

The description of the target language skills provided by the Exam syllabus allows for any task type to be employed together with any type of text on any topic. What is important, though, is that the Syllabus declares the ‘professional dimension’ of the examination which, being very vague by itself, may be suggestive of the difference between ‘professional’ and ‘general’ language skills and the necessity to test both at the Final Examination.

The Examination is obligatory for all graduates and may be considered as internal – defined, designed and administered by the FL Department.

5.2. Content and format of the current Final Language Examination for university graduates

As stated in the Final Examination Syllabus, the Final Language Examination consists of three parts. It must be assumed, because it is not stated otherwise, that these have equal weight:

Exam Task 1: Linguistic theory

Exam Task 2: Listening and summarising (oral summary)

Exam Task 3: Reading and summarising

The format of the Final Language Exam can also be traced through the structure of examination cards (Picture 5.2).

Picture 5.2. An examination card for the Final Language Examination

<p style="text-align: center;">МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Тульский государственный педагогический университет им. Л.Н. Толстого»</p> <p style="text-align: center;">ГОСУДАРСТВЕННЫЙ ЭКЗАМЕН</p> <p style="text-align: center;">по специальности «Иностранный язык (английский)»</p> <p style="text-align: center;">ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 19</p> <p>1. Elaborate on the theoretical problem. The semantic structure of the sentence. 2. Listen to the tape, summarise the content, then enlarge on the topic. 3. Comment on the newspaper article.</p> <p>Билет рассмотрен и утвержден на заседании кафедры английской филологии пр. №4 от 07.12.11 Зав. кафедрой _____ Билет рассмотрен и утвержден на заседании уч. совета ф-та иностранных языков Декан ф-та _____</p>	<p>Final Language Examination (English)</p> <p>Card number</p> <p>Q1 is different for every student</p> <p>Students are given different text according to the examination card number</p>
---	---

All three tasks are performed orally: students take one exam card each and have 60 minutes to plan and prepare their answers to all three questions/tasks. Students are allowed to take notes. The tasks are done in English; Russian is not supposed to be used (see a detailed description of Exam procedure in Part 5.3: Exam administration).

For **Exam Task 1**, a question on **Linguistic theory**, the students are expected to enlarge upon the given issue that lies within Theoretical Phonetics/Phonology, Lexicology, Theoretical Grammar, History of English. Exam Task 1 presents a linguistic topic (Table 5.1) which, as can be seen from the samples below, do not have the form of a question but rather define the area in which students are expected to demonstrate their awareness:

Sample question 1: The Category of Mood (the Indicative, the Conjunctive, the Imperative moods; the meaning of unreality as the common grammatical feature of all verbs expressing Oblique moods varieties; morphological variants and morphological synonyms in the mood system)

Sample question 2: Speech Sounds (classification of vowels and consonants; principles of classification; segmental and suprasegmental speech levels)

Sample question 3: The History of Modern Irregular verbs

Table 5.1. Representation of linguistic subjects in the topics for in Task 1 at the Final Language Exam

Lexicology	The system of English vocabulary
	English affixation
	Synonyms in English
	Conversion in English
	Homonymy in English
	Antonyms in English
Theoretical Grammar	The Noun and its categories
	The Category of Mood
	Parts of Speech
	The Phrase. Principles of classification
	The Categories of Tense, Aspect, Temporal correlation
	The Semantic Structure of the Sentence
	The Category of Voice and Actual Division of the Sentence
History. of E.	The History of Modern Irregular verbs
	The Outline History of the English nominal system
Theoret. Phonetics	Composite Sentences: Complex and Compound sentences
	English Intonation
	Sentence in the Text
	Word Accent in English
	Speech Sounds

(Final examination syllabus developed by the FL faculty, Tula State Pedagogical University, 2010)

The topics, currently 21 in number (Table 5.1), are available before the Examination but students cannot know which one there will be in the Examination card so they get ready for all of them.

As Table 5.1 demonstrates, out of 21 Exam topics for Task 1, 2 are on History of English, 6 on Lexicology, 8 – on Theoretical Grammar and 5 on Theoretical phonetics. This difference in the number of topics on each subject within Task 1 might be explained by difference in the length of the courses: for example, History of English is shorter than Theoretical Grammar in terms of the number of academic hours.

Expected performance: students are expected to answer the given question in English, demonstrating their grasp of the subject-matter, linguistic terminology and, all in all, speaking skills (prepared monologue). The students are supposed either to give examples or comment on (compare, contrast, generalise, etc.) language items given in the Examination card. Exam takers are expected to discuss linguistic issues, but no discussion is expected on how those linguistic issues should be taught at school or any other level. Questions from members of the Examination Board are likely but not obligatory. The number of such additional questions, or the situations when they should be asked are not defined in the Syllabus.

Exam Task 2: Listening and speaking

There is one task type employed in all examination cards (see sample Exam card, p.114): ‘Listen to the tape, summarise the content, then enlarge on the topic’. The texts can be both dialogues (conversations, fragments of radio programmes) and monologues. The length of texts varies from 3 to 5 minutes and the texts are played twice. Texts are different for different students. The only requirement to texts for listening at the Exam, which is presented in the Exam Syllabus, is the topic/theme it covers. The following topics are recommended by the Syllabus:

- | | |
|---------------------------|--------------------------------|
| 1) Appearance | 8) Fashion, clothes |
| 2) Character | 9) Music |
| 3) Family and marriage | 10) Arts |
| 4) House and home | 11) Moscow |
| 5) Free time and holidays | 12) London |
| 6) Sport and exercise | 13) Russia |
| 7) Shopping | 14) English-speaking countries |
- (Final Examination Syllabus, 2010: 38-39)

The list of topics defines the areas in which graduates are expected to communicate. However, there is no detailed description of student expected performance in the Final Examination Syllabus. It can be presumed that the task assesses listening for detailed understanding, so the students are expected to give a detailed summary and comment on the issues the text dwells upon. There are no assessment criteria for students' listening and speaking performance.

The Syllabus does not provide information on how listening texts should be chosen and what sources should be employed. There are also no requirements to the difficulty level of input texts for listening, their length and type.

Exam Task 3: Reading and speaking

There is one task type used in all Examination cards – ‘Comment on the newspaper article’ (see sample Exam card, p.114). The texts are newspaper/magazine articles (from printed or digital sources) of about 500 words (sample texts are presented in Appendix 6). While reading, students are not allowed to use dictionaries or other reference materials. According to the Final Exam Syllabus, the *topics/themes* of texts can be varied within the 14 topics presented in the Threshold level (2001) and broad pedagogical issues concerning upbringing, education, social issues:

- | | |
|--------------------------|--------------------------------|
| 1) Choice of profession | 8) Environment |
| 2) Travelling | 9) City. Living in a big city |
| 3) Food and drink | 10) Health and health service |
| 4) Nature | 11) Courts and trials |
| 5) Mass media | 12) Cinema |
| 6) Hobbies and interests | 13) Being a university student |
| 7) Theatre | 14) Society and values |

(Final Examination Syllabus, 2010: 39)

It can be assumed that the focus of assessment in Task 3 is reading for detailed understanding and prepared monologue because, according to the Syllabus, the students are expected to understand to the full extent the content and implications of the article, give a summary and comment (express their opinion) on the issues, similarly to Task 2. There are no assessment criteria for the Reading and speaking task. There is no indication on what should be assessed in student answers: degree of text understanding, the ability

to present the content of a text, accuracy and/or range of language means. For integrated tasks like the ones under consideration (Listening and Speaking; Reading and Speaking), in addition to assessment criteria, the issue of weighting comes to the fore. The balance between text understanding (listening/ reading) and presenting and discussing its content (speaking) should be made clear.

Similarly to the choice of listening texts for Task 2, no information has been found in the Exam Syllabus on how reading texts should be chosen. There are also no requirements to the difficulty level of texts for reading, their length and type.

The content and format of the Final Language Examination is presented in Table 5.2.

Table 5.2. Final Language Examination: content and format

	Content	Areas of assessment		Text type	Text length**	Task type
		Knowledge	Skills			
Exam Task 1	Linguistic theory <ul style="list-style-type: none"> • Phonology • Grammar • Lexicology • History of Language 	Linguistics Linguistic terminology [no list available]	Prepared monologue (note taking is allowed)			Comment on/ describe how... (see sample questions)
Exam Task 2	Listening and speaking		Listening for detailed understanding Prepared monologue and [possible] answering examiners' questions	dialogues/ conversations monologues	3-5 mins British/ American English	Listen and comment on the issues...
Exam Task 3	Reading and speaking		Reading for detailed understanding Prepared monologue	Newspaper article	~ 500 words	Read and comment on the issues raised in the text

(based on the analysis of the Final Examination Syllabus, 2010)

** As demonstrated by samples of Exam materials; not directly stated in the Exam syllabus

5.3. Administration of the current Final Language Exam for university graduates

Guidelines for Exam administration were sought in the Final Exam Syllabus because this is the only document that defines/describes how the Exam is to be administered. The State Standards issued by the Ministry leave final exam content and administration at university's discretion, so there is no reason to expect this information there. The lower rank documents such as orders issued by the Dean's office mostly aim at technical issues such as appointment of the Examination Board members or allocating students into exam cohorts, but not at regulating Exam administration. The Examination is administered by members of the Examination Board appointed by the Dean's order. The Examination Board consists of FL Faculty staff members with only the Head of the Board being a representative of another university.

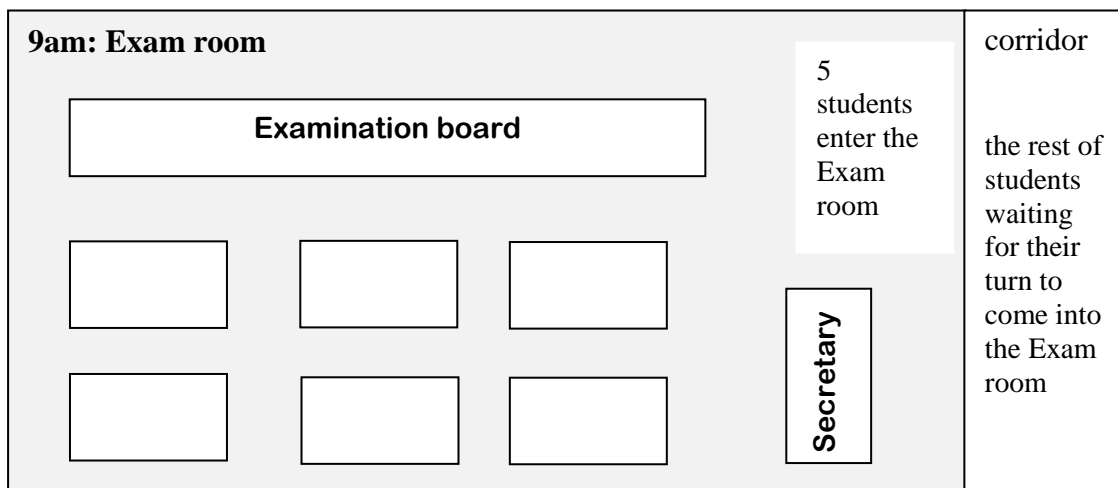
According to the Final Exam Syllabus, the Examination is administered orally (examiners listen to a student's answer to the question in the card), candidates are given 60 minutes each to prepare their answer. They are free to make any notes they need (the stamped sheets of A4 paper are provided). Some students prefer to write exactly what they are going to say, some students write key statements, examples, etc. Only stationery is allowed to be brought to the Examination room.

Before starting their oral answer exam-takers sign the notes that they took while preparing their answers. When they finish answering they leave the notes with the examiners. The notes are not taken into consideration when the final mark is given (e.g. spelling, punctuation, paragraphing and other features of a written piece) but the notes must be neat and accurate, they must be signed by the student and the date must be put. The notes are kept in the Dean's office as an evidence of each student being present at the examination and the mark being given for a 'real answer'. The notes are considered the only proof of student being in the Exam room as no recordings (audio/video) are kept and no electronic/password registration is provided before/at the Examination.

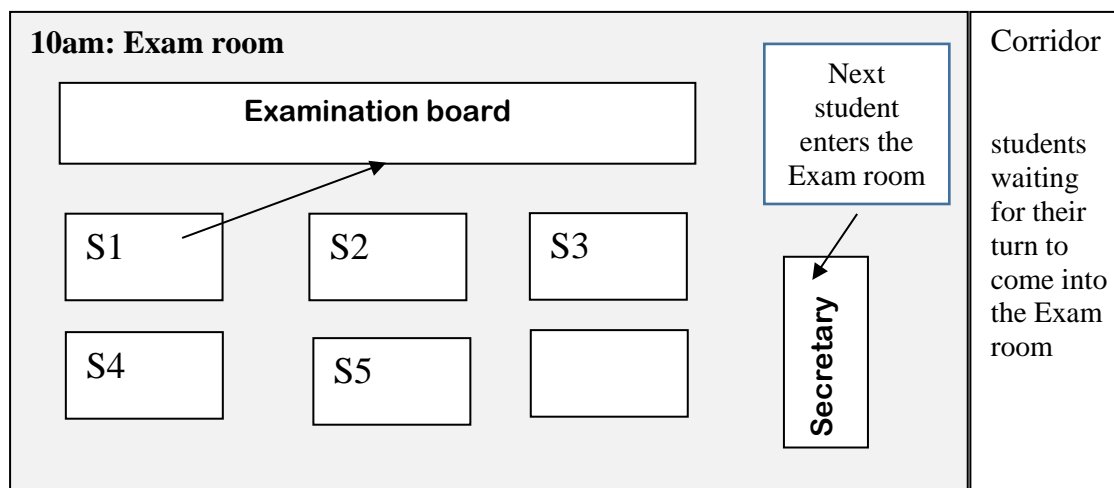
The Examination usually starts at 9am. The students are divided into cohorts (the lists of students in each cohort are usually available a week before the examination). The first five students come into the examination room, take one exam card each and get

ready for their answer. Not all Examination Board members might be present in the Examination room within these 60 minutes – out of five members two may be present. Their function at that time is to observe students in the Exam room getting ready and prevent cheating. Schematically, the Exam procedure may be presented through the diagrammes below (Pictures 5.3a-b).

Picture 5.3a: Examination room at the beginning of the Final Language Exam



Picture 5.3b: Examination room from 10am till the end of the Final Language Exam



1. At 9am the first group of students (usually 5 people) comes in the examination classroom. Each student takes one examination card which has 3 questions.
2. Each student gets a text for reading from the secretary. All texts are numbered and are given to students according to the number of their examination card: Card №20 = Text №20. For listening the procedure is the same: the number of the exam card = the number of the CD track. Each text is played twice.

3. Students sit at the desks (one at each) and start preparing their answers. They are free to choose the one to start with. For all three questions (Linguistic theory, reading and listening) they have 60 minutes.
4. The Examination Board comes in at approximately 10am to start listening to answers.
5. The first student starts answering. Another student (the 6th one) comes into the examination room, takes an exam card and texts.
6. Students answer one by one, in front of other students. Only the last student in the Exam cohort answers with no peers listening to their answer
7. When all students finished answering and left the Exam room, the examination board starts discussing the marks. Usually, each question for every student is discussed separately and the board agrees on the mark.
8. All marks are agreed upon, the Board is ready to announce them
9. The students are invited back to the Exam room for the final marks to be announced. The Chairperson usually comments on strong and weak points (in general). Students are welcome to ask questions if something is not clear. Examiners are supposed to justify their decision on students' final mark, if required.
10. The Examination is over

There is no information in the Final Exam Syllabus or other documents on assessment criteria: what criteria should be applied and if the same set of criteria applies to all three tasks, whether assessment criteria should have the same weight (with Task 1 being quite different from Task 2-3). The only requirement to be met is that the final marks are announced on the day of the Final Examination.

Chapter 5 presents a description of the Final Language Examination based on the documents – State Standards for teacher development issued by the Ministry of Education of Russia (2010), the Final Exam Syllabus (2010) and Final Exam materials samples (revised in 2010) developed by Tula State Pedagogical University. Chapter 7, 9 and 10 present empirical findings obtained from different groups of stakeholders in order to complement the information presented in the documents and get a detailed description of the Final Language Examination, its strengths and weaknesses, as presented earlier in the Research questions (pp.3-4).

Chapter 6

Research Methodology

This research is a case study of the current Final Language examination for language teachers (university graduates) in one of the pedagogical universities in Russia. After Hamilton (2013), the case study is seen as ‘an approach to research that aims to capture the complexity of relationships, beliefs and attitudes within a bounded unit, using different forms of data collection and is likely to explore more than one perspective’ (2013: 10).

This study does not focus on a large research population but rather on small groups of different stakeholders. The research aims to answer **questions** about contexts and practices of the Final Language Examination:

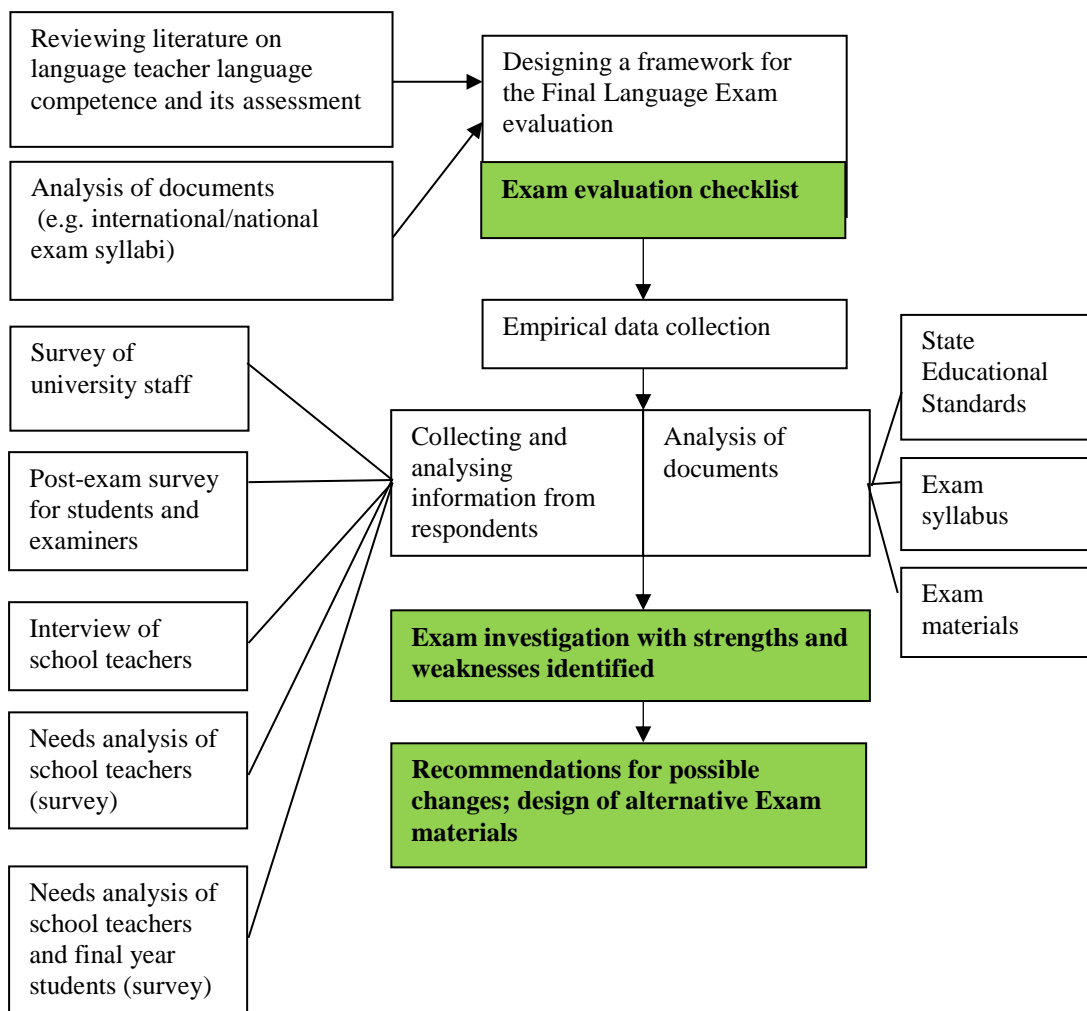
1. What are the procedures for Exam design, piloting and administration as seen by different stakeholders? This includes investigation of:
 - procedures for selecting content and defining the format of the Examination;
 - design and choice of assessment tasks;
 - design and use of assessment criteria.
2. How relevant is the Exam content to the language needs of practising English teachers? What are the language needs of language teachers in Russia?
3. What are the strengths and weaknesses of the current examination? What changes, if any, might be required?
4. What are the possible alternative versions to the current Final Language Examination for language teachers?

The empirical part of this research aimed to collect data from documents and various stakeholders in order to get as complete a picture as possible of the Examination content, format and administration, as well as of Exam design. It is hoped that the data obtained would contribute to detailed Exam evaluation and recommendations for possible changes in the Final Language Examination content, format and administration. The empirical part of the research was performed in several steps, as Figure 6.1 (p.123) demonstrates.

Chapter 6 starts with an overview of the research design – from the development of the Exam Evaluation Checklist, to choice of research methods and design of research instruments (part 6.1). Next, the procedures for empirical data collection (qualitative and

quantitative) and analysis are described (part 6.2). Then ethical issues in research are discussed (part 6.3). The chapter concludes with a discussion of the limitations of this research.

Figure 6.1. Stages of this empirical research³⁷



6.1. Research design

Design of the research instruments was informed by theoretical findings in the area of language testing and test evaluation. The review of literature (Chapter 3, 4) resulted in

³⁷ Analysis of documents is seen as a part of empirical research. Its results are presented in Chapter 5, preceding the Research Methodology chapter. By this means it is hoped to identify gaps in information on Exam design, content and format and provide the necessary context for data collection from various stakeholders.

designing the Final Language Exam evaluation checklist (pp.107-109). The Checklist includes the key parameters for Exam evaluation – from designing Exam tasks to announcing results. These data are essential for drawing conclusions on the current Exam’s validity, reliability, authenticity and practicality and defining Exam strengths and weaknesses.

The data required for investigation of the Exam came from two major sources – documents issued at different levels of the Russian system of education, and stakeholders involved in Exam design, administration and marking. Data collection procedures are presented in Table 6.1.

Table 6.1. Data collection procedures

Type of information obtained	Sources of data	
	documents (Chapter 5)	stakeholders
The process of Examination materials development by university staff including training in administration and materials design		Data from the Faculty of Foreign Languages staff – examiners, exam developers
Description of content, format and administration of the current Examination	State Educational Standards for FL teacher development (Russia) Final Language Examination Syllabus Exam materials samples	Data from the Faculty of Foreign Languages staff – examiners, exam developers
Discussion of Examination characteristics:		
<i>Face validity</i>		Data from school teachers of English as a Foreign Language
<i>Content validity</i>	Analysis of Exam Syllabus; comparison of content of the Syllabus and Exam materials	
<i>Construct validity</i>	Analysis of language teacher language competence (review of publications); comparison of findings with aims and content of the Final examination	
Investigating <i>appropriacy</i> of the format and content of the Final Examination to FL teaching Investigating Exam text and task <i>authenticity</i>		Data from school teachers of English – needs analysis
	Comparing skills under assessment to skills that comprise language teacher language competence Comparing the range of tasks that teachers perform to the range of tasks employed by national and international exams for language teachers	
Exam <i>reliability</i>	Final Exam syllabus analysis 'Chairperson' reports analysis	Data from the Faculty staff (examiners) and exam takers (university graduates)

Investigation of the Final Language Examination involved obtaining information from various sources and employed both qualitative and quantitative methods of data collection. The choice of data collection methods was informed by findings in the area of research design (e.g. Fowler, 1993; Nunan, 1995; Barker, 2005; Wiersma, 2005; Cohen, 2007). Applicability and appropriacy of various data collection methods were considered, using the following parameters (Table 6.2):

- suitability for obtaining qualitative/quantitative data;
- advantages and threats they provide for participants and researchers;
- types/options available (e.g. unstructured/semi-structured/structured interview; open/close-ended self-response items).

Table 6.2. Strengths and weaknesses of empirical data collection methods

<i>Method Options available</i>	<i>Strengths</i>	<i>Weaknesses</i>	<i>Applicability/relevance and restrictions</i>
Quantitative/qualitative data			
<i>Questionnaire</i> e-mail/internet :: pencil and paper questionnaires open-ended:: close-ended questionnaires	Suitable for gathering factual information, data on attitudes, beliefs and experiences (Cohen, 2007) ** Suitable for gathering large scale data ** Generates statistically manipulative data (Nunan, 1995) ** Reliable due to anonymity; encourages great honesty (Cohen, 2007) ** More economical in terms of time (Cohen, 2007)	Percentage of return may be low (Cohen, 2007) ** Possibility of questions being misinterpreted (Fowler, 1993; Barker, 2005) ** No opportunity for respondents to explain their choices (Barker, 2005) ** Open-ended responses are difficult to quantify (Nunan, 1995) ** Can be left blank (Barker, 2005)	Applicable for obtaining data from Department staff (examination materials design, defining exam format and content; exam administration; marking and grading) Suitable for English teacher needs analysis (language skills they apply in and out of the classroom) Suitable for obtaining data from examiners and exam takers on exam administration
<i>Observation</i> Structured:: Unstructured Participant::non-participant Natural settings:: artificial setting (Flick, 1998)	‘What people do may differ from what they say’ (Robson, 2002), so observation provides validity check ** ‘it is advisable to supplement self-report data with observational data’ (Barker, 2005: 3)		Non-applicable (although may yield valuable data) due to ethical issues – may interfere with exam administration and influence the results of the Final (high-stakes) examination

Qualitative data			
<i>Interview</i>			
Informal conversational	A chance to avoid misunderstanding of questions, so fewer sources of error **	Subject to group dynamics (pressure, inferiority) (Barker, 2005) **	Suitable for obtaining attitudes of school teachers on the current Final language examination (its relevance and appropriacy for FL teaching) – allows for additional questions, comments and discussion
Standardized open-ended	Can generate a wide range of responses (Cohen, 2007) **	Low overall reliability (Cohen, 2007) **	
Closed qualitative (Patton, 1980: 206)	Respondents react to each other's contributions so the topic is explored more deeply (Barker, 2005)	Time-consuming	

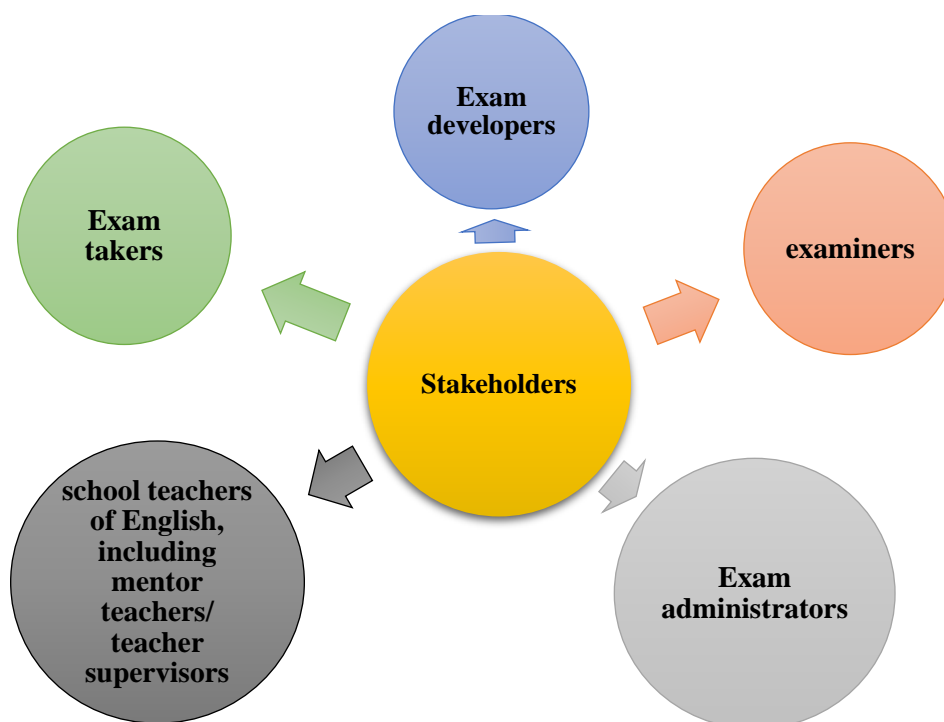
The empirical part of the research aimed at collecting both *qualitative* and *quantitative* data through document analysis, surveys that target different groups of the research population, and semi-structured group interviews concerning issues of Exam materials design, Exam administration, examiners' visions of the Exam focus and adequacy of this focus, as well as other strengths and weaknesses of the Exam.

Data collection was performed according to the principles of triangulation – ‘the use of two or more methods in the study of some aspect of human behavior’ (Cohen, 2007: 141), i.e. studying behaviour (Final Language Examination for language teachers) from more than one standpoint – applying different methods (qualitative and quantitative) to the same object of study. Apart from applying several methods, the research involved various groups of respondents: surveys and interviews dialogued with different stakeholders – from final year university students (Exam takers) to school teachers of English and Exam developers at university level (Figure 6.2).

Surveys were based on self-administered questionnaires, one of which is on-line. The questionnaires consisted of both close-ended and open-ended items with the latter requiring short answers to avoid ‘incomplete, vague and difficult to code responses’ (Fowler, 1993: 100). The questionnaires aimed at gathering factual information with a small degree of attitudinal information from a relatively large population (Wiersma, 2005; Cohen, 2007). *The interview*, by contrast, targeted a much smaller group of English teachers and aimed to obtain their opinions on and attitudes to the current Final Language Examination content and format. Other key issues of research design –

designing and piloting instrumentation, sampling, data analysis – are discussed further in part 6.2.

Figure 6.2: Stakeholders involved in data collection in this research



6.2. Empirical data collection and analysis

As stated in part 6.1, data collection was performed through analysis of documents (presented in Chapter 5), surveys and interviews.

Survey 1, administered in English, involved university lecturers who take part in the Final Language Examination as examiners, materials developers and administrators. The major purpose of Survey 1 was to obtain data on the Exam design (content and task selection, task design and piloting) and administration (assessment criteria employed, how marking is administered). Survey 1 also aimed to investigate the question of which features the Exam designers and examiners see as problematic and in need of change, and which they see as Exam advantages.

Survey 2, also administered in English, was a post-exam survey for examiners and final year students (Exam takers). Its purpose was to get information on Exam administration – the factor essential for Exam evaluation that could not be fully traced through documents.

Survey 3 was a Needs analysis of English teachers in Russia (Tula region). It aimed at getting data on the knowledge and communicative skills teachers employ in their work, the tasks they perform in and out of the classroom, and the types of texts they deal with. These data were used in defining appropriacy of texts and tasks in the current Final Language Examination.

Survey 4 was a web-survey for final year university students (future teachers of English), and was a variation of Needs analysis in Survey 3. It aimed at getting data on tasks students performed during their teaching practice, what knowledge and skills these tasks involved and how easy/difficult students found these tasks.

The **interview** dialogued with school teachers of English as a foreign language in the Tula region. The purpose of the interview was to involve the participants in reflection on and analysis of the current Final Language Examination tasks from the perspective of language teachers and language users, i.e. to express their opinions on exam task appropriacy and relevance for the FL teacher job.

6.2.1. Quantitative data

Survey 1 involved twenty respondents who were involved in the design and administration of the Exam from Tula State Pedagogical University, in different roles (examiner/administrator/task designer) and with different levels of experience in Final examinations (from 1 to 10 years). No sampling procedures were planned as the aim was to involve everybody taking part in the Final Language Examination. Twenty-one requests to fill out the questionnaire were sent by e-mail, and twenty responses were obtained which made the response rate 95%.

Table 6.3: Final Exam Survey.

Cross-tabulation: respondents' age*experience in the Exam

		experience				Total
		less than 2 years	2-5 years	6-10 years	more than 10 years	
respondent's age	39 or under	1	5	2	1	9
	40-60	0	0	2	6	8
	61 or over	0	0	0	3	3
Total		1	5	4	10	20

Table 6.4: Final Exam Survey.

Cross-tabulation: respondents' experience*role in the Exam

		Role in the Final Language Examination			Total
		materials developer	assessor	examiner	
respondent's experience	less than 2 years	1	0	1	
	2-5 years	4	4	5	
	6-10 years	3	4	4	
	more than 10 years	9	9	10	
Total		17	17	20	20

In literature on language testing (e.g. Alderson, 1995; Heaton, 1995; Hughes, 2003; McNamara, 2000; Bachman&Palmer, 2010) there is a division of roles into examiners and assessors. This is only required for oral exams like the one under study, or oral parts of other examinations, including standardized ones. The examiner's role is seen as administration of oral tasks – introducing a situation, asking questions at an interview, and maintaining a conversation with the exam taker. Examiners are not supposed to take notes, write down exam takers' mistakes, etc. The assessors' role, on the contrary, is to take detailed notes and/or make audio-/video recordings in order to make marking as objective as possible. Assessors do not interact with exam takers and often sit so that exam takers do not face them.

As Tables 6.3 and 6.4 demonstrate, respondents with different levels of experience – from quite limited (within 2 years), to rather extensive (10 years and more) – perform the same functions: developing Exam materials and working as examiners and assessors at the Final Examination which is quite common for internal assessment like the Exam under consideration. Although there is no clearly stated division of responsibilities, a tendency to involve more experienced staff in performing all Exam functions can be traced.

The survey employed a specially designed questionnaire (Appendix 7) available in English in electronic and printed versions. There was no difference between the versions; the participants were free to choose any of the two since this was done for their convenience only. Four respondents asked for a printed copy. The rest of the respondents preferred a Microsoft Word file. The questionnaire was sent by e-mail to the Head of the English Department at Tula State Pedagogical University and then was sent back by the respondents, saved on the hard disk and coded so that respondents' identity could not be traced.

The content of the questionnaire was informed by the outcomes of the literature review (Chapter 3, 4) and the Exam evaluation checklist (pp.107-109). The questionnaire aimed to cover the key areas of language testing in general and language testing for language teachers in particular:

- focus and purpose of Final Language Examination – knowledge areas, communicative skills in general and teacher English; principles of content selection and representation;
- Exam administration procedures, including assessment criteria and their weighting, and marking procedures;
- Final Language Exam materials design, moderation, pre-testing.

The questionnaire consisted of 35 questions:

- 9 multiple-choice (multiple answers);
- 16 multiple-choice (one answer, including yes/no/not sure);
- 6 open-ended short answer questions;
- 4 scales (agree-disagree; always-never) designed on the basis of the Likert attitude scales.

Issues to be addressed	Number of item	Item type
Test materials design:		
Task selection	Q1-5	M.choice + open-ended
Text (reading, listening) selection	Q6-7	M.choice + open-ended
Moderation and pre-testing	Q8-11	M.choice
Who is involved in task design	Q12	M.choice
Exam staff training	Q13-16	M.choice
Exam format and content	Q17-18	M.choice + attitude scale
Assessment procedures	Q19-23	M.choice + open-ended
Exam administration	Q24-29	M.choice + open-ended
Exam strengths and threats	Q30-33	M.choice + attitude scale
Potential changes in exam format/content/both	Q34-35	M.choice

Before the questionnaire was launched, it was piloted on a limited sample (5 participants), all of whom were lecturers at the Department of English and English Philology of Tula State Pedagogical University. Participants with varying experience in the Final Language Exam and performing different roles (examiner/materials developer/both) were offered the questionnaire together with a short piloting checklist to fill out.

Table 6.5: Piloting checklist for Survey 1

	<i>yes</i>	<i>no</i>
Do you understand the survey's objective?	5	
Do you feel comfortable answering questions?	5	
Is the wording clear?	2	3
If not, what is not quite clear?	2 people felt uncomfortable about 'inter-rater' and 'intra-rater reliability'; they and one more person misunderstood the role of 'exam administrator'	
Are there any items which are too long or difficult to answer?		5
Are there any items which produce irritation, embarrassment or confusion?		5
If yes, which ones?	-	
Have any important issues been overlooked?	1	4
If yes, which ones?	<i>It was suggested that the department be specified</i>	
N=5		

Appendix 7 presents the final version of the questionnaire, after piloting changes were introduced.

After the responses were obtained, they were coded and analysed – quantitatively and qualitatively. An SPSS database was created for quantitative responses (Figure 6.3, p.134), whereas responses to open-ended questions were grouped and analysed in accordance with the procedures suggested in the literature.

After Nunan (1995: 145-147), the following steps were taken:

- verbal responses (Q4, 5, 6, 7, 8, 19, 22, 23) were written out and organized in a table according to the number of the question and the code for the respondent (Table 6.6, p.134);
- colour coding was applied to highlight repeated patterns (words/phrases) so that categories could be generated from the responses. In the sample from Table 6.6 the presented responses are short (as planned and expected), so that often the repeated pattern was equal to the category generated. The whole set of categorized qualitative data is presented in Appendix 14.

Figure 6.3: SPSS database for the Final Exam Questionnaire

	age	mats_dev	exam_admin	assessor	other	experience	staff	sch_l_teach	others	Stands	syllabus	spr
1	39 or under	yes	no	yes		2-5 years	yes	no		yes	yes	
2	40-60	yes	yes	yes		more than ...	yes	no		yes	yes	
3	39 or under	yes	no	yes		more than ...	yes	no		yes	yes	
4	39 or under	no	no	yes		2-5 years	yes	no		yes	yes	
5	61 or over	no	no	yes		more than ...	yes	no		yes	yes	
6	61 or over	yes	yes	yes		more than ...	yes	no		yes	yes	
7	39 or under	yes	no	yes		2-5 years	yes	no		yes	yes	
8	39 or under	yes	no	yes		2-5 years	yes	no		yes	yes	
9	39 or under	yes	no	no		less than 2...	yes	no		yes	yes	
10	39 or under	yes	no	yes		6-10 years	yes	no		yes	yes	
11	40-60	yes	no	no		more than ...	yes	no		no	yes	
12	39 or under	yes	no	yes	syllabus wr...	6-10 years	yes	no		yes	yes	
13	40-60	no	no	yes		6-10 years	yes	no		yes	yes	
14	40-60	yes	no	yes		more than ...	yes	no		yes	yes	
15	40-60	yes	no	yes		more than ...	yes	no		yes	yes	
16	39 or under	yes	no	no		2-5 years	yes	no		yes	yes	

Table 6.6: Example of categorizing qualitative responses in Survey 1

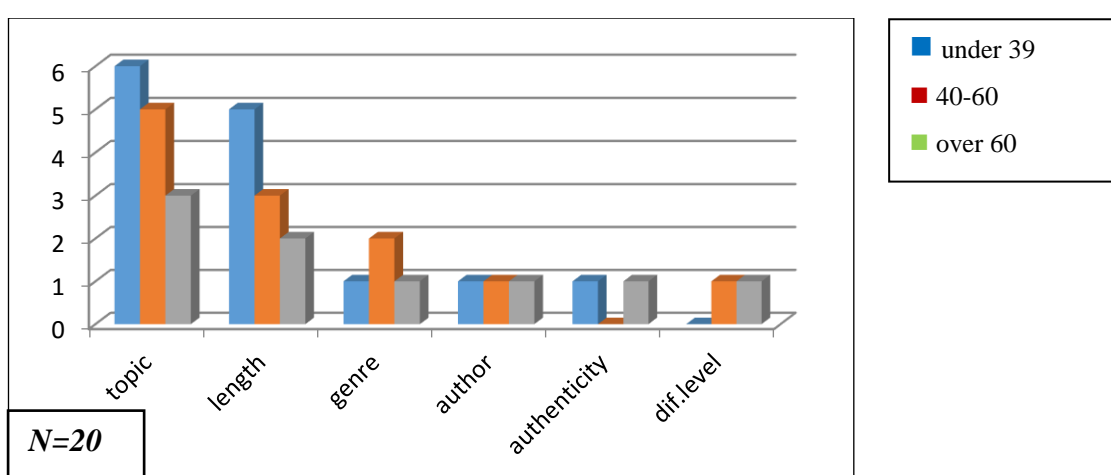
Q7 Criteria for selection of reading texts Categories: TOPIC LENGTH GENRE AUTHOR AUTHENTICITY DIFFICULTY	<i>Exam_staff_1</i>	Authentic, approx 1 page, based on topics studied
	<i>Exam_staff_2</i>	
	<i>Exam_staff_3</i>	Topic, length, genre
	<i>Exam_staff_4</i>	The situation ... is the same as with listening
	<i>Exam_staff_5</i>	Topic, genre, length, author, authenticity
	<i>Exam_staff_6</i>	In accordance with the list of conversational topics (studied during the course) and the level of difficulty (corresponding to C2 – Proficient User)
	<i>Exam_staff_7</i>	Topics, length, text type, when published (if an article)
	<i>Exam_staff_8</i>	Topics, length, author, date of publication
	<i>Exam_staff_9</i>	I just follow what has been done before me
	<i>Exam_staff_10</i>	Topic, length, text type (mostly newspaper articles)
	<i>Exam_staff_11</i>	Topics and genre
	<i>Exam_staff_12</i>	Topics
	<i>Exam_staff_13</i>	See listening
	<i>Exam_staff_14</i>	
	<i>Exam_staff_15</i>	Topics or issues they dwell upon, length, source??? (author's punctuation)
	<i>Exam_staff_16</i>	
	<i>Exam_staff_17</i>	Topics, author (when we used fiction), length, difficulty
	<i>Exam_staff_18</i>	Topic and length of the text
	<i>Exam_staff_19</i>	Topic, probably the length of the text, genre
	<i>Exam_staff_20</i>	Just as listening – according to the topics + length of the text

- the categories were introduced into SPSS and cross-tabulated to see if there was some difference in opinions depending on respondents' age, role and experience of exams (see Table 6.7).

Table 6.7: A sample presentation of participant responses to Q7, Survey 1: cross-tabulation age*criteria for choosing Exam texts

Age	Criteria for reading text selection					
	topic	length	genre	author	authenticity	difficulty
under 39 <i>N=9</i>	6	5	1	1	1	--
40-60 <i>N=8</i>	5	3	2	1	-	1
over 60 <i>N=3</i>	3	2	1	1	1	1
Total <i>N=20</i>	14	10	4	3	2	2

Figure 6.4: A sample of visual presentation of qualitative responses to Q7, Survey 1



The complete set of statistical data obtained through Survey 1 is presented in Chapter 7 and further discussed in Chapter 8.

Survey 2 was a post-exam survey for examiners and students (Exam takers). This small-scale survey involved five examiners and 11 students at the Final language examination in February 2013 at Tula State Pedagogical University. The survey was based on two specially designed questionnaires – Questionnaire 2A for examiners, and Questionnaire 2B for students (Appendix 8A, 8B).

Survey 2A for examiners and Survey 2B for Exam takers addressed similar issues of Exam administration. The content of both questionnaires was informed by the outcomes of the literature review, namely the section that dealt with language test reliability and the ways it can be ensured. Surveys 2A and 2B dealt with timing issues

– time for preparation, time spent answering; examiner intervention and its effects (as seen by examiners and Exam takers); possibility of using reference materials – the issues which contribute (or do not contribute) to equality of conditions for all exam takers. The data obtained from Surveys 2A and 2B complemented the data on Exam administration procedures, marking and grading from Survey 1. Therefore, a section of Questionnaire 1, and Questionnaires 2A and 2B were centred on similar issues – Exam administration, assessment and marking.

Apart from the content, the difference between Survey 1 and 2 lay in the time they were administered and their purpose:

- Survey 1 concentrated on a wider range of issues and mostly dealt with general, repetitive situations, whereas Survey 2A was mostly about what was going on for a particular cohort of students on a particular day in terms of Exam administration, conditions Exam takers and examiners were in, difficulties (if any) they faced and how such difficulties were dealt with.
- Survey 2B addressed the same issues as 2A through the eyes of another group of stakeholders – students who took the Exam.

Table 6.8: Differences between the content of Survey 2A and 2B

Issue to consider	Survey 2A (examiners)	Survey 2B (exam takers)
Exam administration		
• number of examiners	✓	✓
• intervention in student answers	✓	✓
• results of intervention		✓
• examiners' attitude		✓
• if examiner behavior is specified	✓	
Assessment procedures	✓	
Clarity of assessment criteria	✓	✓
Duration of the examination	✓	✓
Exam content		✓

The *essential requirement* to be met was that the questionnaires were filled out on the day of the Examination, after the Exam takers finished their answers and before the final marks were announced. This was done with the assistance of the Head of the IT Group at the English Department of Tula State Pedagogical University. The time at which questionnaires were offered to Exam takers was chosen on purpose, so that the

announced results of the Final Examination could not influence the Exam takers' perceptions or attitudes to what was going on in the Examination room.

Questionnaires 2A and 2B consisted of close-ended and open-ended items with close-ended ones prevailing. The responses were divided into quantitative and qualitative data. Two SPSS databases were created for quantitative responses – one for each survey with variables defined by the content of the questionnaires and research questions. Qualitative responses (separately for Survey 2A and 2B) were grouped according to the questions they were provided for, were analysed and categorized and were further transferred into qualitative data as Table 6.9 demonstrates.

Table 6.9: Categorizing responses to open-ended questions in Survey 2

<p>Any commentaries about examination</p> <p>Categories CRITERIA DIFFICULT TO APPLY</p> <p>UNCLEAR ASSESSMENT FOCUS</p> <p>ASSESSMENT INSTRUMENTS TO BE RECONSIDERED</p>	P_EX_E_1 ³⁸	<p>There is a list of criteria which is difficult to apply. I know what accuracy means and what fluency means this is not helpful. The most important seems to be content as they answer a linguistic question. But what if they make language mistakes?</p> <p>Something should be reconsidered about criteria</p>
	P_EX_E_2	<p>I do not think it is possible to assess the answer to the first question and the second question together and announce one mark. It's not clear what we are assessing. Even if it is knowledge – what kind of knowledge is it? The criteria can only be applied to question 2 and even there it is not clear how to apply them. I am not happy with this</p>
	P_EX_E_3	<p>I am not happy with the new format – there is only one task when they answer a linguistic question. It is not clear at all how the criteria apply to this task. What shall I assess? Content or accuracy or fluency or everything? No, it's not clear</p>
	P_EX_E_4	<p>I believe that our State exam doesn't really check students' language skills. It rather checks their ability to memorize a lot of theoretical material and their experience as school teachers</p>

The complete set of statistical data obtained through Surveys 2A-B is presented in Chapter 7 and further discussed in Chapter 8.

Survey 3 Needs analysis of teachers of English as a Foreign Language aimed to:

- identify knowledge and communicative skills (in English as a FL) that English teachers employ in their everyday professional life. Division into listening, reading,

³⁸ Stands for Post-Exam-questionnaire, Examiner 1/2/3/4

speaking and writing follows the Common European Framework of Reference (2001); the European Profile for Language Teacher Education (2004); some language examinations for teachers (e.g. LPATE) and research projects in teacher English (e.g. Skuja&Mee, 1997; Sešek, 2007);

- compile a taxonomy of tasks which teachers deal with;
- provide data for discussing the validity of the current Final Language Examination for language teachers;
- provide a rationale for suggested changes in the Exam content and format.

The survey employed a specially designed questionnaire (Appendix 9) as a means of data collection. The questionnaire was available in printed and electronic versions. The versions are absolutely identical and respondents could choose the one they felt more comfortable with, e.g. at the seminar for language teachers (November 15, 2012), 18 participants were offered printed questionnaires whereas other respondents preferred electronic versions as they found them easier to deal with.

The questionnaire consisted of five questions all of which presented opinion/attitude scales based on the Likert scale. The respondents were asked to mark communicative skills in English as always/often/seldom/never employed in their professional life.

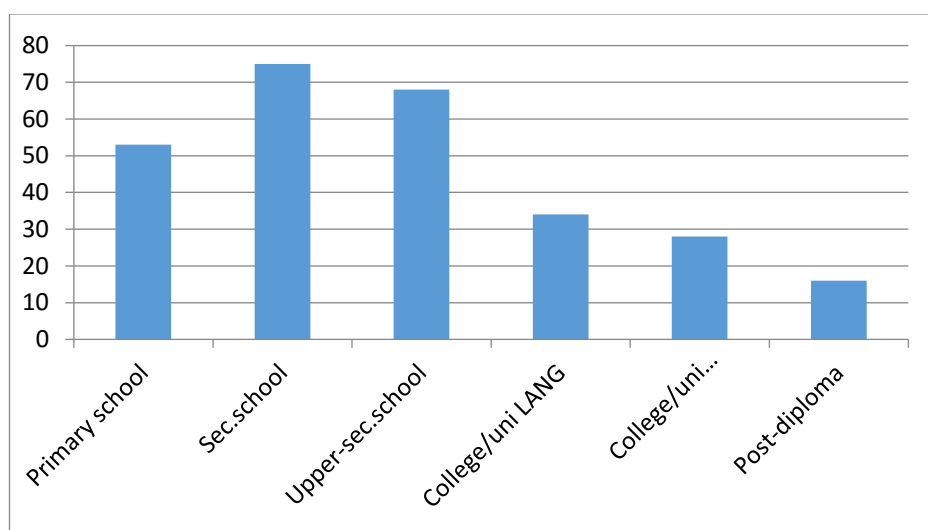
Issues to be addressed	Number of item	Item type
How often are listening skills employed? What skills are they?	Q1	Scale (never::every day)
How often are reading skills employed? What skills are they? What texts do teachers deal with?	Q2	Scale (never::every day)
How often are speaking skills employed? What skills are they?	Q3	Scale (never::every day)
How often are writing skills employed? What skills are they?	Q4	Scale (never::every day)
How confident are English teachers in different areas of English?	Q5	Scale (extremely unconfident::very confident)

The content of the questionnaire was informed, first of all, by findings in the area of language teacher language competence - communicative skills (listening, reading, speaking, Classroom Language, writing) and knowledge areas that teachers employ in and out of the classroom. As stated previously, the purpose of this questionnaire was to suggest a taxonomy of skills that are required by English teachers in the Tula region.

This was compared to the list of skills that are currently assessed at the Final Language Examination and served as a rationale for possible changes in the Exam.

The research population comprised school teachers of English, teachers of English (lecturers and senior lecturers) from Tula State University, Tula State Pedagogical University, Tula State University of Military Engineering, Tula Branch of Police Academy of Russian Federation, and Tula Academy of Tourism and Catering Industry. The questionnaire was sent to 90 e-mail addresses of school principals (the list of schools is available from the website of Tula Ministry of Education) and Heads of English departments at universities who were asked to cascade the message down to their teachers and lecturers. One hundred and seven people responded, with the estimated size of the whole target group being 500 people. The sample was self-selected, only those teachers participating who wanted to answer the questionnaire. More than 60% of respondents were school teachers at different levels - primary, secondary, upper-secondary. The remaining 40% were teachers of English from universities (see below).

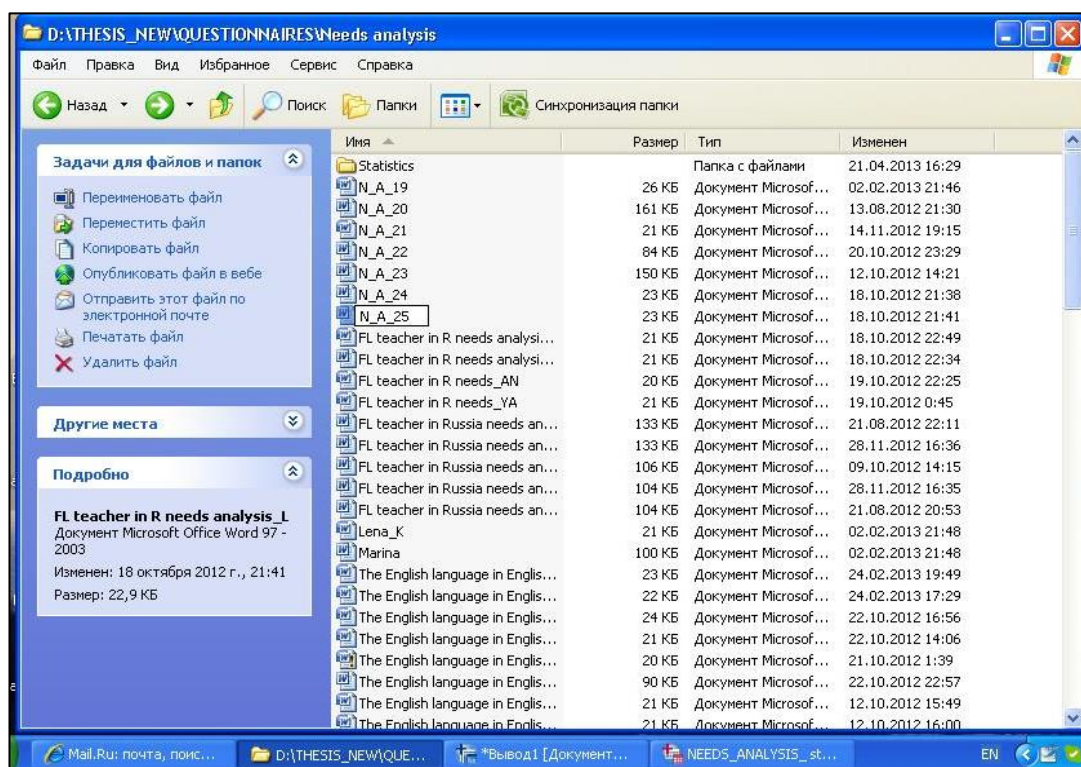
Figure 6.5: Research population in Survey 3: levels the respondents teach at³⁹



³⁹ The total number of responses is more than 107 because most of the respondents teach at more than one level: for example, it is typical for a teacher to teach at primary, secondary and upper-secondary school, or for a college lecturer to teach at university and at in-service post-diploma courses

Data collection was performed according to the purpose of the research in general and the purpose of the survey described above. Eighteen questionnaires were filled out in the printed form, whereas the remaining 89 respondents preferred the electronic version and sent their responses by e-mail. All responses were coded so that no respondent's identity could be traced (Picture 6.1).

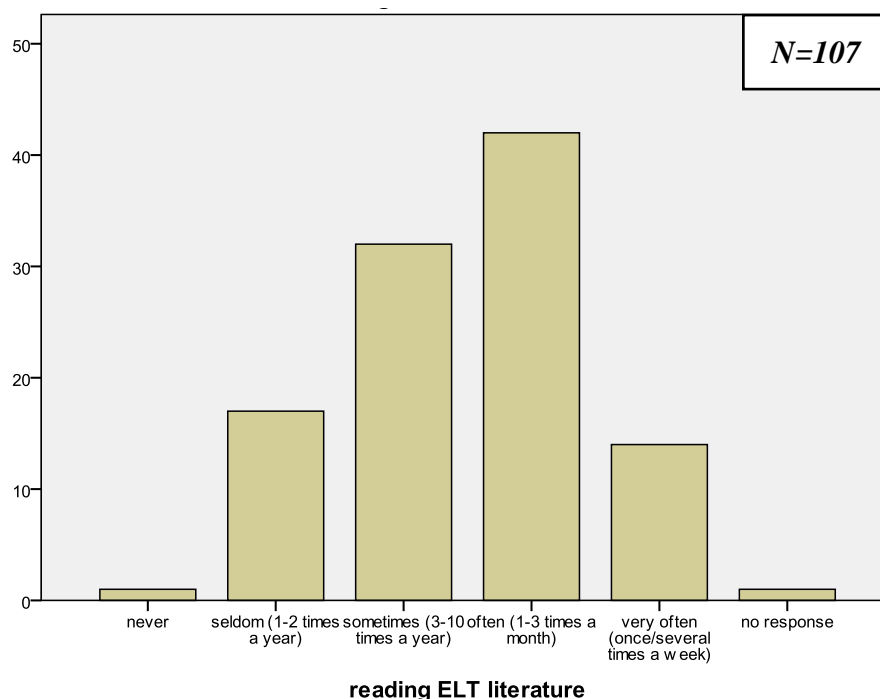
Picture 6.1: Survey 3: coding responses sent by e-mail⁴⁰



All the responses are close-ended, and they were categorized in SPSS. With SPSS, data were generated according to the research questions. The major interest was in how often teachers of English employed this or that communicative skill in and out of the classroom. In this chapter, Figure 6.6 and Tables 6.10, and 6.11 present samples of how statistics were generated for illustrative purposes only. The results obtained through Survey 3 are presented in Chapter 9 and discussed in Chapter 10.

⁴⁰ N_A stands for Needs Analysis. Coding of electronic responses started with N_A_19 because the printed versions of the questionnaire were coded from N_A_1 to N_A_18

Figure 6.6: Statistics from Survey 3: How often teachers read ELT literature



The diagramme in Figure 6.6 was produced irrespectively of respondents’ age, work experience or school type. It demonstrates that the options most frequently chosen by respondents who answered the question were ‘sometimes’ and ‘often’. Then, the data was cross-tabulated according to respondents’ work experience.

Table 6.10: A sample of data analysis for Survey 3:
Cross-tabulation experience*reading ELT literature

		reading ELT literature						Total
		never	seldom (1-2 times a year)	sometim es (3-10 times a year)	often (1-3 times a month)	very often (once/several times a week)	no response	
work experience	5 years or less	0	4	6	9	2	0	21
	6-10 years	0	1	9	8	2	0	20
	11-20 years	0	6	9	14	5	0	34
	more than 20 years	1	6	8	10	5	1	31
Total		1	17	32	41	14	1	106

The data presented in the table above were distributed in a manner similar to that presented in the diagram (Figure 6.6) – the most frequent options for all four groups (less than 5 years, 6-10 years, 11-20 years, more than 20 years) still being ‘sometimes’

and ‘often’, with a similar curve on the diagram. The same procedure was applied to school types resulting in cross-tabulation of school type*reading ELT literature.

Table 6.11: A sample of data analysis for Survey 3:
Cross-tabulation of school type*reading ELT literature

	reading ELT literature						Total
	never	seldom (1-2 times a year)	sometimes (3-10 times a year)	often (1-3 times a month)	very often (once/ several times a week)	no response	
secondary school	0	10	21	31	12	1	75
college/university	1	6	10	13	4	0	34

As can be seen from Figure 6.6 and Tables 6.10 and 6.11, a similar trend can be observed in both datasets – the options chosen more frequently are ‘sometimes’ and ‘often’, irrespectively of the respondents’ working experience or type of school they work in. Thus, the data from Needs Analysis were generated and presented only according to the skills and sub-skills that teachers employ, without subdivision of the responses into categories according to school type or respondents’ experience (Chapter 9). In Chapter 10 these data are also discussed as a whole.

Survey 4: Needs analysis of final year students

The purpose, structure and content of this survey were similar to those of Survey 3 – Needs analysis of English teachers. The data sought in Survey 4 dealt with language skills that students used at their Teaching Practices, how often they needed those skills and which of those skills they found easy and which seemed challenging. Survey 4 aimed at collecting data from final year university students (future teachers of English) who had accomplished all rounds of teaching practice and were approaching their Final Language Examination. The final year students were considered an important group of stakeholders with only minimal teaching experience but relevant visions of their professional development. This group of ‘Needs analysis’ respondents received a different questionnaire, for the following reasons:

- the activities students are involved in during their Teaching Practice are similar but not identical to activities of practising teachers, so a difference in applying language skills can be observed for these two groups;
- the list of skills to include in the questionnaire should agree with the list prescribed by the Teaching Practice syllabus;
- the gradation in scales is the same as in Needs analysis for teachers – from ‘never’ to ‘very often’, but their meanings are different for these two categories of respondents due to different amount of professional experience. For example, using a skill ‘sometimes’ for experienced teachers meant 3-10 times a year whilst for trainee teachers ‘sometimes’ implied ‘5-6 times during their teaching practice’ (Appendix 9, 10).

The survey was administered on-line through Survey Monkey with the link <http://www.surveymonkey.com/s/KCY39MC> sent to the Head of the English department of Tula State Pedagogical University. The on-line form was preferred by respondents as it was quickest and most convenient for them being available 24/7 from any computer or mobile device. The on-line questionnaire consisted of eight questions: six close-ended (3 multiple-choice and 3 scales) and two open-ended, requiring short answers.

Picture 6.2: On-line survey: screenshot – multiple choice (multiple answers)

The screenshot shows a web browser window with the SurveyMonkey interface. The question is labeled 'B2' and has several control buttons: 'Редактировать вопрос', 'Добавить логику вопроса', 'Переместить', 'Копировать', and 'Удалить'. The question text is: '*2. In what type(s) of school(s) did you have your teaching practice? Choose as many as applicable.' Below the question are four radio button options: 'state comprehensive school', 'state language school/lyceum/gymnasium', 'private school', and 'other (please specify)'. There is an empty text input field below the 'other' option.

Picture 6.3: On-line survey: screenshot – scale

	Never or once	Seldom (2-4 times)	Sometimes (5-6 times)	Often (nearly every week)	Very often (nearly every day)
Listening to pupils performing in the classroom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Observing English lessons	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Listening to recordings to coursebooks	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Watching (taking part) in ELT seminars	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading teacher's books/resource packs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading ELT literature, including magazines	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The responses obtained were collected and analysed by Survey Monkey. Verbal responses to Q6 and Q7 were collected and categorised manually to be used further in Chapter 9 and Chapter 10.

6.2.2. Qualitative data

Qualitative data were obtained through a series of interviews with 11 respondents involved. Three group interviews took place (Group 1 = 4; Group 3 = 2; Group 4 = 4 participants), and one individual asked to be by herself. A structured type of interview was chosen, for which the exact wording and sequence of questions were defined in advance, so all interviewees were asked the same questions in the same order (e.g. Patton, 1980, Nunan, 1995). According to Wiersma (2005), and Cohen&Manion (2007), structured (or standardized open-ended) interviews have some important strengths with a minimal number of weaknesses (Table 6.12).

Table 6.12: Strengths and weaknesses of structured interviews

Strengths	Weaknesses
Respondents answer the same questions, thus increasing comparability of responses; data are complete for each person on the topics addressed in the interview. Reduces interviewer bias when several interviewers are used. Permits decision-makers to see and review the instrumentation used in the evaluation. Facilitates organisation and analysis of data.	Little flexibility in relating the interview to particular individuals and circumstances; standardized wording of questions may constrain and limit naturalness and relevance of questions and answers.

(Cohen & Manion, 2007: 353)

The interview framework was designed and handouts (samples of current Final Language Examination tasks) prepared in advance (Appendix 11). The purpose of the interview was to investigate teachers' opinions on the content and format of the current Final Language Examination for language teachers, i.e. appropriacy and relevance of Exam tasks to tasks teachers of English perform regularly in and out of the language classroom⁴¹. These data were further used to:

- add to and clarify the tentative description of language teacher language competence (Chapter 3);
- provide some evidence for content validity evaluation for the current form of the Exam (Chapter 10);
- provide a rationale for possible changes in Exam content and format (Chapter 12).

The structure and content of the interview framework were defined by its purpose. Each interview comprised two parts: the introduction and ice-breaker, and the main part.

Table 6.13: Structure of the interview for teachers of English

	Aim	No of quest.	Question type(s)	
Intro	to introduce participants;	6	Open-ended, short answer M. choice (multiple answers)	
	to get information on participants' teaching context (work load, age group, level they work at, school type, routines and activities)	1		
Main part	to obtain teachers' opinions on relevance and appropriacy of exam tasks	5	Open-ended, extended answer M. choice (one answer)	
		3		
	to obtain teachers' visions of possible changes in the exam		1	M. choice (multiple answers)
			2	M. choice (one answer)
		1	Open-ended, extended answer	

⁴¹ Tasks performed by teachers out of the classroom include a range of tasks from lesson planning to browsing the Internet for classroom activities and reading ELT literature. Tasks performed in the classroom involve conducting classes, maintaining discipline, explaining language items, etc.

The eleven interviewees were teachers of English of different ages, with different levels of experience and working at different levels at different types of schools - comprehensive, language schools; private language schools (Tables 6.14, 6.15, 6.16a-d). Participation in the interviews was strictly voluntary. School teachers of English were addressed at the in-service seminar for teachers (Tula, November 2015) and also when Survey 4 (Needs Analysis) was administered. Fifteen teachers from Tula and regional towns wished to participate, with the final number declining to 11. Four potential participants had to withdraw due to health and other issues.

Table 6.14: Teacher interview: participants' age

		Number of people	Percentage of total number of participants
Interviewees' age	under 39	10	90.9%
	40-60	1	9.1%
	Total	11	100%

Table 6.15: Teacher interview: participants' teaching experience

		Number of people	Percentage of total number of participants
Interviewees' experience	1.5 years	1	9.1%
	3 years	1	9.1%
	5 years	3	27.3%
	7 years	1	9.1%
	8 years	1	9.1%
	9 years	2	18.2%
	11 years	1	9.1%
	18 years	1	9.1%
	Total	11	100%

Four groups were organized: two groups of four people, one group of two people and one individual interview (pp.144-146). The group forming principles were experience and school type, i.e. people with the same level of experience but from different school types were invited to the same group. This provided some common ground, stimulated discussion and, at the same time, excluded 'mentor-mentee' relationships which are often observed in groups of teachers of different ages, as well as helping to avoid or minimize conformity pressure that would give more weight to more experienced/prestigious members (e.g. Barker, 2005).

Table 6.16a. Interview 1: participants

<i>Experience</i>	<i>School type</i>
9 years	comprehensive (town)
9 years	comprehensive (city)
5 years	private language school
8 years	private tutor

Table 6.16b. Interview 2: participants

<i>Experience</i>	<i>School type</i>
19 years	comprehensive

Participants of *Interview 1* were all graduates of Tula State Pedagogical University. Three of them entered Tula Teacher Training College #1 first and obtained their teacher qualification there. Then they went to the University to obtain Higher Professional Education. One participant was a graduate of the 5-year teacher training programme at university. Three participants held their teacher qualification with distinction.

Three participants currently work with city pupils (Tula) – at comprehensive schools or privately. One participant works in a comprehensive school in her native town, teaching the full range of age groups.

The participant of *Interview 2* graduated in 1995 from Tula State Pedagogical University and has worked in a comprehensive school in Tula since then. She works with a wide range of age groups – from 7 (primary) to 16 (upper-secondary school) and also as a private tutor.

Table 6.16c. Interview 3: participants

<i>Experience</i>	<i>School type</i>
7 years	comprehensive + college
10 years	private language school

Table 6.16d. Interview 4: participants

<i>Experience</i>	<i>School type</i>
5 years	comprehensive (town)
1.5 years	private language school
3 years	private language school (children)
5 years	private language school (adults)

Both participants of *Interview 3* were graduates of the College-University model (described in Chapter 1). They graduated from Tula Teacher Training College #1 and then finished the University course. One held her diploma with distinction. One is currently employed as a secondary school teacher in Tula, whilst the other works in a private language school in Moscow, working mostly with adults.

Three of four participants of *Interview 4* were tutors in one of the most popular private language schools in Tula. The school selects its staff very carefully and invests considerable resources in staff in-service training, so the requirements for the staff are much higher than in comprehensive schools. One participant works in a comprehensive school, with all age groups, in her native town 100km away from Tula. She is the only English teacher in the school.

Participants of Interview 1 and 3 had similar amounts of experience and could have been interviewed as one group. Still, the decision was made to split those teachers into two groups because, first, a group of six interviewees would have been difficult to manage so as to give them all equal chances of expressing themselves, and, second, participants of Interview 3 knew each other and preferred to stay in one group/pair. All interviews were audio-recorded and transcribed (Appendix 12). Interview 2 (individual) was conducted in Russian at the teacher's request, so it was transcribed in Russian and then translated into English. Responses to questions from the introductory part of the interview and close-ended questions from the main part were introduced into an SPSS database.

Picture 6.4: Teacher interview: introducing quantitative data in SPSS

	age_group	experience	comprehen	gymnasium	priv_language	other_type	primary	secondary	upper_sec...	adults	load	les
1	under 39	9 years	yes	no	no		yes	yes	yes	no	28 hours	
2	under 39	9 years	yes	no	no		yes	yes	yes	no	31 hours	
3	under 39	8 years	yes	yes	yes		no	yes	no	no	15 hours	no
4	under 39	5 years	yes	yes	yes		yes	yes	yes	yes	11 hours	
5	40-60	19 years	yes	no	no		yes	yes	yes	no	31 hours	
6	under 39	7 years	yes	no	no	college	yes	yes	yes	yes	24 hours	
7	under 39	11 years	yes	no	yes		no	no	yes	yes	8 hours an...	
8	under 39	5 years	yes	no	no		yes	yes	yes	no	23 hours	
9	under 39	1,5 years	no	no	yes		yes	yes	yes	yes	25 hours	
10	under 39	3 years	no	no	yes		yes	yes	yes	yes	19 hours	
11	under 39	5 years	no	no	yes		no	yes	no	yes		
12												

Elements of content analysis (e.g. Nunan, 1995; Cohen&Manion, 2007: 474-475) were applied to analysis of open-ended responses, i.e.:

- data from all four interviews were grouped according to the questions discussed, e.g. everything that was said about the Linguistic part of the examination was put

together in a table in a Microsoft Word file. The same procedure was applied to discussion of other parts of the exam, etc.

- units of analysis were defined – mostly phrases and sentences;
- repeated patterns were first colour-coded, then utterances were shortened to phrases. Next, shortened utterances were grouped according to colour, their meaning was generalized and categories were introduced. Some categories emerged from the data itself (e.g. *irrelevance to teacher’s job, unnecessary complexity of knowledge*), whereas some were influenced by the literature review and the exam evaluation checklist (e.g. *knowledge about language :: ability to teach*);
- to ensure validity of categorization employed in this research, inter-rater coding was administered, i.e. the same coding task being performed by different people with different backgrounds (co-coding). The interview scripts were first organized into a table according to the questions that the respondents had answered. Then the tables were given to three people who did not know each other:
 - a non-native speaker with a TESOL qualification;
 - a native speaker with no TESOL qualification;
 - a native speaker with a TESOL qualification.

They were asked to categorize the same chunks of data using colour coding and coming up with categories of their choice. Then the results were compared with each other and with the initial categorization, resulting in the final list of categories (Appendix 13: Tables 1-4). As a last step, main features of the situation were identified, and inferences were drawn; the results were presented in Chapter 9 and discussed in Chapter 10.

Table 6.17 presents a sample of the final version of coding which sums up responses from all people involved.

Table 6.17: Teacher interview: categorization of open-ended responses (a sample)

Extracts from interviews	CATEGORY	Interview number
LINGUISTIC PART		
The teacher must know the subject she teaches and that’s why they [materials] are absolutely appropriate. It must be there	IMPORTANCE OF LINGUISTIC KNOWLEDGE IMPORTANCE OF THIS PART OF EXAM	1 (group)
Because we can be asked questions from our students concerning...why this or that phenomenon is used in the language and it is	IMPORTANCE OF LINGUISTIC KNOWLEDGE	1

important ... to be able to answer... so these questions are appropriate		
I am not sure that teachers can use so deep knowledge working at school. Of course, teachers should know a lot but I think such tasks are more for university teachers, for those who want to devote their lives ... well, I don't know... who want to become linguists. Or those who want to teach Linguistics. This is too deep for school, pupils do not need it. When they [pupils] ask, do you tell them in that detail?	<p>UNNECESSARY COMPLEXITY OF KNOWLEDGE</p> <p>IMPORTANCE OF LINGUISTIC KNOWLEDGE</p>	1
Oh, it's so difficult to say. [...] You know this and you know that but then you explain the difference between Present Simple and Present Continuous and you see that they do not understand. And what you (points at everybody) know about P Simple and P Continuous does not help. I would change something, but I don't know what. These questions should be there. I mean, teachers must know theory. But it should be less complicated... or more practical.	<p>DIFFERENCE BETWEEN KNOWLEDGE ABOUT LANGUAGE AND ABILITY TO TEACH</p> <p>UNNECESSARY COMPLEXITY OF KNOWLEDGE</p> <p>IMPORTANCE OF THIS PART OF EXAM</p> <p>IMPORTANCE OF LINGUISTIC KNOWLEDGE</p>	1
Well... no doubt we need it. From the linguistic point of view. But from the point of view of teaching practice... it's not too good. Well, it's difficult to say. I do not remember the time I needed it. It is a bit too deep, too detailed. We do need it at a simpler level. We need it, but not in that detail. I would put it in another way: we do need it and NOT at a simple level. But not at the exam... it should be a separate stage and should take place earlier – not at the State exam. I would not change anything. I would administer this part before [emphasizes] the State exam. I do not know how it can be done, but it should be done.	<p>IMPORTANCE OF LINGUISTIC KNOWLEDGE</p> <p>DIFFERENCE BETWEEN KNOWLEDGE ABOUT LANGUAGE AND ABILITY TO TEACH</p> <p>NEVER APPLIED WHAT I LEARNED</p> <p>UNNECESSARY COMPLEXITY OF KNOWLEDGE</p> <p>EXAM FORMAT/CONTENT SHOULD BE RECONSIDERED</p>	2 (individual)
I think it's rather appropriate. I think a teacher must have certain background knowledge of language theory. It's (meaning the exam questions) probably too much ... it's more than needed at a comprehensive school. But we must be able to answer any tricky question like "Why it is so?" But this is ... probably too deep. A multiple choice test? What about a test? Matching could be employed to check terminology.	<p>IMPORTANCE OF LINGUISTIC KNOWLEDGE</p> <p>UNNECESSARY COMPLEXITY OF KNOWLEDGE</p> <p>EXAM FORMAT/CONTENT SHOULD BE RECONSIDERED</p>	3 (group)
I don't know. I work mainly with adults. Corporate clients, you know. It's rather difficult to think about usefulness from the position of a school English teacher. I think all this knowledge is important. We must know these things. But I do not remember I used this knowledge.	<p>IMPORTANCE OF LINGUISTIC KNOWLEDGE</p> <p>NEVER APPLIED WHAT I LEARNED</p>	3

<p>I have to compare with Russian. When I teach Grammar. And I explain difference. But I never go very deep... they do not need it.</p> <p>I am sorry... but I have to say I do not remember everything. You know what I mean? If you ask me one of these questions I would answer... I think... but not in detail.</p> <p>I would somehow reduce this part. Probably... less questions... or questions should be smaller.</p>	<p>UNNECESSARY COMPLEXITY OF KNOWLEDGE</p> <p>EXAM FORMAT/CONTENT SHOULD BE RECONSIDERED</p>	
<p>I think this part is appropriate. Because I don't believe that practice is possible without theory.</p> <p>In fact, it's based on it.</p> <p>I still think it's just right.</p> <p>And it should be there.</p> <p>I would leave everything as it is.</p>	<p>IMPORTANCE OF LINGUISTIC KNOWLEDGE</p> <p>IMPORTANCE OF THIS PART OF EXAM</p>	4 (group)

The procedures described above resulted in a multi-faceted analysis of the current practice, with strengths and weaknesses of the Final Language Examination in question identified and possible alternative ways of developing it suggested.

Table 6.18 demonstrates how the research instruments were used to collect data on the Research Questions. Each research question is addressed by several data collection instruments, although some surveys (e.g. Survey 3, 4) target more limited specific areas.

Table 6.18: Relation of research instruments to research questions

<i>Research question</i>	<i>Instruments employed to collect data</i>
<p>1. What are the procedures for Exam design, piloting and administration?</p> <ul style="list-style-type: none"> • content selection • design and choice of tasks and input • design and use of assessment criteria <ul style="list-style-type: none"> • Exam administration 	<p><i>Survey 1: Q1-7, Q12</i></p> <p><i>Survey 1: Q8-11, Q12-16</i></p> <p><i>Survey 1: Q19-23</i></p> <p><i>Survey 2: questions about assessment criteria</i></p> <p><i>Survey 1: Q24-29</i></p>
<p>2. How relevant is the Exam content to the language needs of English teachers? What are those needs?</p>	<p><i>Survey 1: Q18</i></p> <p><i>Teacher interviews</i></p> <p><i>Survey 3</i></p> <p><i>Survey 4</i></p>
<p>3. What are the strengths and weaknesses of the current Final Language Examination? What changes, if any, are required?</p>	<p><i>Survey 1: Q30-33</i></p> <p><i>Survey 1: Q34-35</i></p>
<p>4. What are possible alternative versions of the Final Language Examination?</p>	<p><i>Survey 1: Q34-35</i></p> <p><i>Teacher interviews</i></p>

Table 6.19: Empirical data collection: summary

Research method	Issues	Instruments	Question types used	Respondents involved	Number of respondents	When data was collected
Analysis of documents	Exam format and content Expected student performance	Checklist (exam evaluation profile)				November 2011 - April 2013
Survey 1	exam materials design, moderation, piloting; exam administration; problems as seen by those involved; possible areas of change	Questionnaire 1 (Final Exam Questionnaire)	multiple choice (multiple answers/one answer), open-ended questions; scales	exam materials developers, examiners, administrators from FL department	20; no sampling – everyone was involved	December 2012 – February 2013
Survey 2	exam administration, application of assessment criteria, inter-rater and intra-rater reliability; marking	Questionnaire 2 (Post-Exam questionnaire)	multiple choice (multiple answers/one answer) ; open-ended questions	Examiners (quest 2a) Final year students - exam takers – (quest 2b)	5 examiners 11 students (1 cohort) ; no sampling – everyone was involved	February 18, 2013 (the day of the Exam)
Survey 3	skills teachers employ; their confidence in different areas of Teacher English	Questionnaire 3 (FL Teacher needs analysis)	multiple choice (multiple answers/one answer); scales	Teachers of English (schools, colleges, unis, private language schools) in Tula region	107; random sampling (school teachers)	December 2012 – March 2013
Survey 4	Same as for Survey 3 but for beginner teachers with very limited experience	Questionnaire 4 (Final Year student needs analysis)	multiple choice (multiple answers/one answer); scales	Final year students of Tula State Pedagogical University	11; random sampling	January – April 2013
Semi-structured interview	Appropriacy and relevance of current exam tasks to the job of a FL teacher	Interview framework; exam materials samples	Multiple choice; open-ended questions	Teachers of English in Tula region	11; represent. sampling	January 2013

6.3. Ethical issues of research

As this research deals with obtaining data from people of different ages (from 21 to over 65) and groups (school teachers, students, lecturers and administrators) ethical issues must be addressed. After Cohen&Manion (2007), and also ethical guidelines published by the British Educational Research Association (2004), ethical issues are dealt with by means of:

- openness (explaining the purpose and planned outcomes of the research in cover letters for questionnaires);
- providing anonymity and confidentiality of participants and data. This was explained in the cover letters to questionnaires and consent forms for interviews;
- providing an opportunity to withdraw from participation in the research at any stage without explaining reasons;
- providing the participants with the processed data from the survey;
- appealing to people's experience and expertise and not 'checking what they know' or treating their answers as right or wrong, i.e.:
 - avoiding specialist terminology
 - avoiding questions like 'enumerate all you know about...'/ 'give the definition of this term'.
 - avoiding evaluative judgements or commentaries in interviews
 - emphasizing how important each opinion is for research purposes
- in focus groups, not involving people in (former) teacher-student/mentor-mentee relationships in one group;
- in focus groups, providing an opportunity to choose between English and Russian as an interview language;
- providing options like 'other'/'more' so that participants have freedom of expression;
- asking participants to express opinions on facts/things but not on specific people (e.g. evaluation of materials but not of their developers' skills).

Anonymity, after Cohen&Manion (2007: 64), is understood as inability of the researcher to identify the participant from the given information and, through this feature, provision

of participants' privacy. This study involved information on participants' age, teaching experience and level at which they work, while simultaneously not revealing people's names and other sensitive information. Confidentiality, as suggested by Cohen&Manion (2007: 65), implies that there is no public access to facts and opinions revealed by particular respondents in surveys and interviews. To avoid the opposite situation, in which no obtained data could be reported due to either anonymity or confidentiality, the responses were coded, no full names were presented, and any references to educational institutions or people were omitted.

Data collection caused no disruption of the teaching process at university. The questionnaires were filled in at times suitable for the respondents and did not require the researcher's presence. At initial stages of research design, observation was considered as a method of collecting data on Final Language Exam administration. Later, due to ethical issues and possible disruption in administration of a high-stake examination, the possibility of observation was reconsidered and developed into Post-exam survey 2 (Appendix 8A, 8B). Questionnaires for Survey 2 were offered to students after they had finished their oral answers. It was made clear that students' opinions would not be available to examiners or in any other way influence students' final grades.

In an attempt to reduce any negative impact of the research on its participants, the purpose of research was highlighted and reiterated throughout data collection. The participants were reminded that the purpose of the research was multi-faceted evaluation of current Exam practices rather than evaluation of staff's decisions or outcomes of Faculty of Foreign Languages work.

6.4. Limitations of research

In this research the most serious concern is absence of a clear definition in the literature of language teacher language competence. Although there is no clear common consensus, a tentative working definition was developed from a thorough analysis and comparison of opinions to be found in publications (Chapter 3).

Another concern is lack of transparency in describing the aims and content of FL teacher development in Russia, and absence of clarity over the roles played by different

educational bodies in Russia and abroad (e.g. the Bologna agreement). This, in the end, resulted in difficulties in accessing resources. To overcome this limitation, every effort was made to involve stakeholders from different institutions and with different statuses and views on the Examination in question.

Participation in surveys and interviews was strictly voluntary, so it can be presumed that only the 'good' representatives wished to participate – those working in good schools, who attend teacher events and have access to resources. However, there is no statistical or documentary evidence of this. The surveys involved teachers from different types of schools, with different levels of experience, from Moscow, Tula and smaller towns in Tula region. For this research it is presumed that the sample is representative enough, and with a different sample the results would be approximately the same.

The last but not least significant limitation is that the Final Language Examination was evaluated in one university only, as a case study of final assessment at Tula State Pedagogical University (Russia, Tula region). Although the situation in Tula is typical of many other universities, no statistical or other data were provided in support of this statement. This research is viewed as an initial step in evaluation of Final language examinations for novice teachers of English. A key outcome of this study, apart from Exam evaluation, is the design of research instruments (4 questionnaires, including one on-line, and an interview framework) that, due to time constraints, could be applied at first only in a limited context like the one under study. The designed Exam evaluation framework can be transferred into a wider context, which would provide more extensive and reliable data.

Chapter 7

Findings on Research Question 1: design and administration of the current Final Language Examination for teachers of English as a Foreign Language in Russia

This chapter presents the data obtained from 2 groups of stakeholders – Exam developers at university level, examiners and Exam takers. The data collected through Survey 1 and Survey 2a, b contributes to getting a detailed picture of:

- Examination materials development:
 - criteria for task and text selection;
 - procedures for materials development, including moderation and pre-testing (piloting), design of specifications;
 - criteria for appointment of materials developers;
- Exam administration:
 - assessment criteria, marking and grading;
 - timing issues, specifications of examiner behaviour.

The data collection was performed in November 2012 – April 2013 in Russia through specially designed questionnaires administered in pencil-and-paper and electronic forms. *The first group* of respondents involved university staff (Faculty of Foreign Languages) – 20 Exam developers and examiners (as the survey demonstrated, there is a little or no division of responsibilities at the Final Examination and all the roles – materials development, assessment, administration – are performed by the same people). The questionnaire is presented in Appendix 7. *The second cohort* of research participants included examiners (5 people) and Exam takers (11 final year students) who were asked to fill out post-exam questionnaires (Appendix 8A, 8B). The major purpose of the survey was to get more detailed information on Exam administration – timing, examiner behavior, reference materials allowed, etc. – immediately after the Examination was finished.

Data collection methodology and the data obtained define the structure of this chapter. It first presents the data according to its sources/respondents and then develops into comparison and analysis of results with its strengths and weaknesses identified.

7.1. Survey of the Faculty of Foreign Languages staff

As presented in Chapter 6: Research Methodology, the survey was based on a specially designed questionnaire (Appendix 7) for the Department of Foreign Languages staff. It aimed at specifying as many issues as possible on Exam development, administration, content, format because the existing documents – State Educational Standards for teacher development and Final Examination syllabus – present either insufficient or quite vague information (Chapter 5: Description of the current Final Language Examination for future teachers of English as a Foreign Language in Russia).

The Questionnaire was subdivided into five parts, each dealing with an issue of Exam development or administration:

- Part I: Examination materials design (Q1-11)
- Part II: Staff selection and training for the Examination (Q12-15)
- Part III: Examination content and format (Q17-23)
- Part IV: Exam administration (Q24-29)
- Part V: Respondents' opinion of the Examination (Q30-35)

7.1.1 Examination materials design

This part of the questionnaire aims at obtaining information on examination materials development at the Faculty of Foreign Languages (Department of English) – how decisions are made and who is involved in the process of defining Exam format and task design.

The first question (multiple choice) was about **people** (staff or non-staff members) **involved in** the process of **Examination materials design**. Table 7.1 presents the results obtained.

Table 7.1: People involved in Final Language Exam materials development

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
FL Department teaching staff at university	20	100%
School teachers of English	0	0
Others	1 (department administration)	5%

As can be seen from the table above, the respondents⁴² – Exam developers and examiners - demonstrated unanimity of responses: everybody (100%) marked the option ‘FL Department teaching staff’ as materials developers with only one person adding Department management (Heads of departments, Dean and/or Vice Dean). It can be explained by the fact that the staff performing managerial roles perform other roles – conducting classes and materials development, so any member of management is definitely treated as Department staff with some additional functions to perform. None of respondents stated that ‘outsiders’, i.e. non-university staff like, for example, school teachers of English, are involved in Examination materials development.

A question closely connected with the previous one deals with **professionals involved in choosing examination task types**. Table 7.2 presents the responses obtained:

Table 7.2: Responses to Q3: *Who is involved in choosing Final Exam task types?*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Faculty management	19	95%
Lecturers and other teaching staff	17	85%
School teachers	0	0%
Others (Faculty Council)	1	5%

The responses clearly show that school teachers and other ‘outsiders’, i.e. non-faculty staff, are not involved in the process, the decisions are made at the level of the university department. Although the respondents’ answers leave no doubt about the people involved, it remains a bit unclear why the option ‘lecturers and other staff’ is not chosen by everybody. One of possible reasons may be a chance that ‘other staff’ option was in some way misleading and could be interpreted as ‘all lecturers of the department making the decision’.

The next question deals with **the resources** to which Examination materials designers have access and the documents they are supposed to follow.

Table 7.3: Resources employed in the Final Language Exam materials design

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
State Educational Standards for FL teacher development	19	95%

⁴² Hereafter in part 7.1 ‘respondents’ include Final Exam materials designers, examiners and Exam administrators – all members of staff of the Faculty of Foreign Languages

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Final Examination syllabus	20	100%
Final Exam specifications	3	15%
Exam materials from previous years	18	90%
Recordings of student performance	1	5%
Other sources (internet resources)	1	5%

As the above demonstrates, State Educational Standards, Final Examination Syllabus and materials from previous years were chosen by almost everybody – 95%, 100% and 90% respectively. It is interesting that three people (15%) chose the option ‘Final Exam specifications’. Exam specifications do not seem to exist in the studied context because in the Russian education system exam specifications are not considered to be an essential document for test design. These are widely discussed in the literature on language testing and understood as ‘the official statement about what the test tests and how it tests it’ (Alderson, 1995: 9). Exam specifications are seen as a detailed document describing test purpose, test taker, test structure, target language situation, text types and their sources; language skills under assessment, language elements to be tested; number of items in each section; test methods and assessment criteria (e.g. Alderson, 1995; Heaton, 1995; H.D. Brown, 2004; Bachman & Palmer, 2010). Exam specifications are barely mentioned in any Russian official documents dealing with assessment and only became essential for the National Language Examination for school-leavers in the year of 2000 (Solovova, 2011, 2014; <http://ege.edu.ru/>, retrieved on November 28, 2014). The function of specifications at all other levels of education in Russia is usually performed by an examination syllabus which does define skills and topic areas under assessment but still the information there is not as detailed as it usually is in specifications. Nevertheless, 3 out of 20 respondents to the questionnaire chose ‘Exam specifications’ as a source in Exam materials development. ‘Recordings of student performance’ was the option chosen by only one respondent. It can be explained by a possible misinterpretation of ‘recordings’ which might have been treated as examiners’ records/notes or the notes that students take while preparing their answer in the examination room. No tape recordings of oral answers have ever been done as it is not a standard procedure at oral examinations at any level in Russia.

Another key area in materials design is **choice of Examination tasks and text types** for Listening and Reading sections. These choices can be partially traced through

sample tasks presented in the Final Examination Syllabus (Chapter 5, pp.114-115). Nevertheless, the better part of issues remains vague. So, survey questions Q4-7 aimed at deeper understanding of the task and text selection process.

Q4 of the Final Exam Survey was a combination of two sub-questions: a close-ended question about clarity of task selection criteria (if any) and an open-ended question on the exact guidelines (if any) task designers are supposed to follow.

The responses to the close-ended question were cross-tabulated according to respondents' experience, as presented in Table 7.4.

Table 7.4: Responses to Q4: *Are the criteria for choosing task types clearly stated in the Exam syllabus/other document?*

		Clarity of criteria for choosing Exam tasks: responses			Total
		yes, clearly laid out	just mentioned	not even mentioned	
Respondents' experience	less than 2 years	0	0	1	1
	2-5 years	0	2	3	5
	6-10 years	0	0	4	4
	more than 10 years	1	0	9	10
Total (out of 20)		1	2	17	20

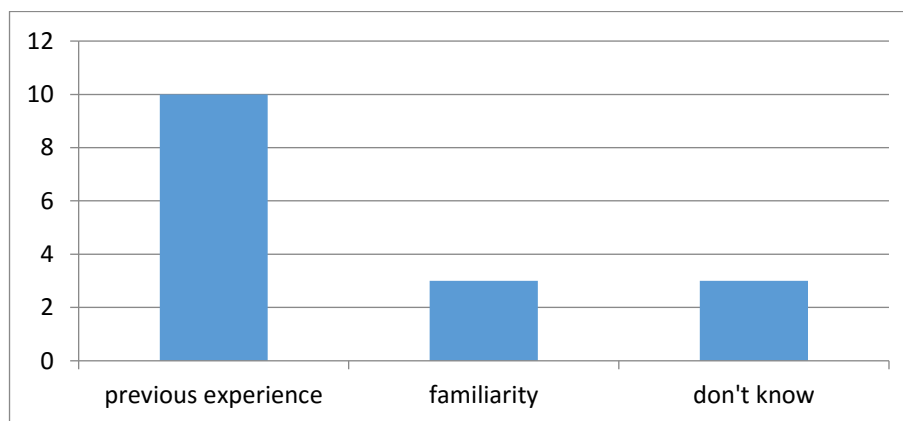
As Table 7.4 demonstrates, 17 out of 20 respondents with different levels of experience stated that the criteria for task selection are 'not even mentioned' in the Final Exam Syllabus or any other document, whereas 1 person with more than 10 years of experience said the criteria were clearly laid out. Two people said that the criteria were just mentioned, meaning, probably, the list of topics the texts should be about.

The answers to the open-ended part of the question (Table 7.5, Figure 7.1) demonstrate that often task designers tend to follow procedures which were set up many years ago or employ tasks that students are familiar with, i.e. those used in the course of studies and for progress assessment. On the one hand, familiarity of task types is considered one of the key requirements to language tests; on the other hand, under the studied circumstances, using particular task types without specifying all possible options might lead to convenience choices and sticking to the same tasks year after year without realizing why this is the only task type worth using.

Table 7.5: Criteria for choosing task types (open-ended responses to Q4: *Are the criteria for choosing task types clearly stated in the Exam syllabus/other document? If yes, what are they?*)

Respondent	Open-ended responses
Exam_staff_1	Guidelines are given by administration. The criteria should cover all or the most important aspects of students' performance in language studies and theoretical subjects ⁴³
Exam_staff_2	⁴⁴
Exam_staff_3	It has always been like that, so we use what proved to be effective
Exam_staff_4	I really don't know (author's emphasis)
Exam_staff_5	The types which have already been chosen
Exam_staff_6	Cambridge ESOL Examinations Standards ⁴⁵
Exam_staff_7	Use what was chosen many years ago
Exam_staff_8	We use them because they were chosen many years ago
Exam_staff_9	??? (as answered by the respondent)
Exam_staff_10	Don't know, just use what proved to be OK
Exam_staff_11	
Exam_staff_12	What's been employed in the previous years
Exam_staff_13	Task types which are used for progress assessment
Exam_staff_14	We use what proved to be good
Exam_staff_15	Tasks which everybody is familiar with
Exam_staff_16	⁴⁶
Exam_staff_17	We use tasks which are used in all years, including exams
Exam_staff_18	We use materials from the previous years
Exam_staff_19	I think we repeat the same year after year
Exam_staff_20	What was used previously, or experience of other universities

Figure 7.1: Criteria for choosing task types (statistical summary of open-ended responses)



⁴³ The response provided does not make the respondent's opinion clear

⁴⁴ No response provided

⁴⁵ The respondent mentions Cambridge ESOL examinations without referring to any particular exam. No references to international language examinations have been found in the Final language exam documents, nor can it be traced through exam format, content or administration

The responses to the follow up question (Q5) about some criteria being more important than others can be viewed as a logical continuation of the task selection issue: 14 out of 19 respondents said they were not sure if some criteria had priority:

Table 7.6: Responses to **Q5**: *Are there any criteria of task selection which are more important than others?*

		Some task selection criteria more important than others			Total
		yes	no	don't know	
experience	less than 2 years	0	0	1	1
	2-5 years	1	0	3	4
	6-10 years	0	1	3	4
	more than 10 years	1	2	7	10
Total (out of 20)		2	3	14	19

Q6-7 of the Final Exam Survey dealt with **text** selection for listening and reading tasks and revealed a similar problem: texts should be chosen according to some guidelines and those guidelines are either vague or not presented at all.

Table 7.7: Responses to Q6-7: *Are the criteria for choosing **listening /reading texts** presented in the Exam Syllabus/other document?*

		Criteria for choosing listening texts			Total
		clearly described	loosely described	not given	
experience	less than 2 years	0	0	1	1
	2-5 years	1	1	3	5
	6-10 years	0	2	2	4
	more than 10 years	1	2	7	10
Total (out of 20)		2	5	13	20
		Criteria for choosing reading texts			
experience	less than 2 years	0	0	1	1
	2-5 years	1	1	3	5
	6-10 years	0	2	2	4
	more than 10 years	1	3	6	10
Total (out of 20)		2	6	12	20

Open-ended responses about text selection criteria were analysed (Appendix 14) with the results presented in diagrammes. Figures 7.2a-b (statistical summary of open-ended responses) demonstrate that for selection of listening and reading texts the key guidelines are topic and length.

Figure 7.2a: Criteria for choosing **listening** texts

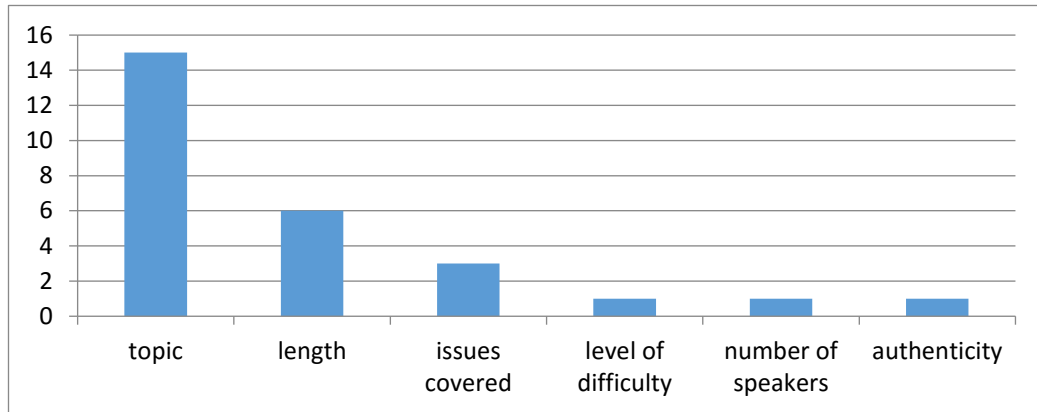
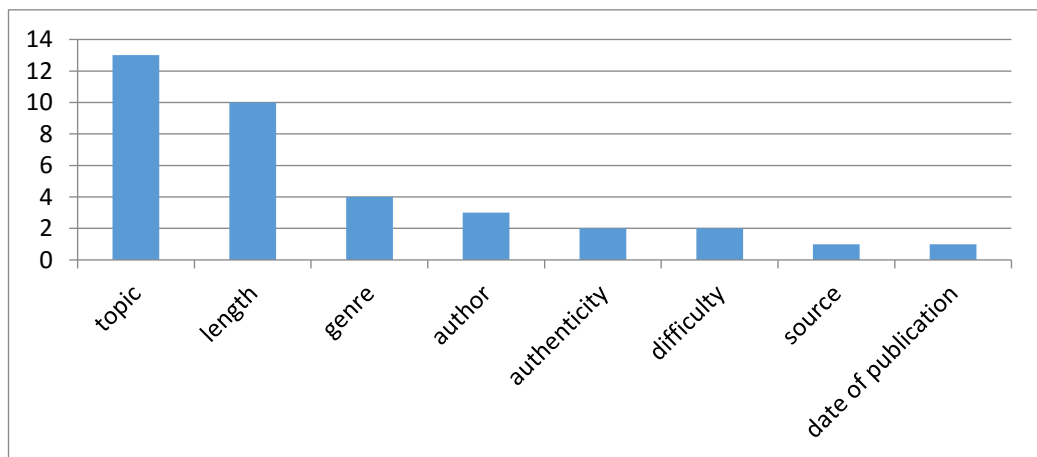


Figure 7.2b: Criteria for choosing **reading** texts



As presented in Chapter 5, the process of examination materials design is not described or even mapped in any document under consideration – either at the Federal or local level. Materials design is a responsibility delegated to the Department staff and there is some evidence (for example, responses to the questions about choices of tasks and texts) to suppose that it follows the procedures employed for years. It might be considered an advantage/strength but the responses obtained from the staff involved in materials design demonstrate some lack of coordination and understanding of the process of materials design and why it is supposed to be in this particular way.

A part of examination materials design is their moderation and piloting. As there is no indication in the documents under review on moderation of Final Exam materials taking place at the FL Department, this question was investigated in Survey 1.

Table 7.8: Responses to Q8: *Are exam tasks moderated (scrutinized by several staff members before they become exam tasks)?*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
yes, always	2	10%
sometimes	8	40%
no	10	50%
Total (out of 20)	20	100%

As the table above presents, two respondents – Exam designers and examiners - said the tasks were always moderated whereas 10 people chose the option ‘never’ and 8 respondents said that items were moderated ‘sometimes’. It means that no indication has been found on a commonly accepted and documented procedure of materials design in general and their moderation in particular, so decisions about item moderation might be made spontaneously.

Consequent to this spread of opinions, there is quite a predictable difference in views about people involved in moderation: 14 out of 20 respondents stated that moderation was performed by the Faculty staff and administration. Six people omitted the question, which in some way, may be viewed as logical, because, answering the previous question (Q8), 10 people said there was no moderation at all. Similarly to the responses to Q8, the difference in opinions observed in Q9 might testify to the fact that there is no commonly accepted and clearly defined procedure of materials design.

Table 7.9: Responses to Q9: *If materials/items are moderated, who is involved in the process?*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Faculty staff	9	45%
School teachers	0	0%
Faculty administration	5	25%
Others	0	0%
No response provided	6	30%

Apart from getting information on moderation of Final Exam materials, Survey 1 asked the respondents about possible materials piloting/trialling. Responses obtained for Q11 demonstrate that Exam tasks (whether they undergo moderation or not) are never trialled:

Table 7.10: Responses to Q11: *Are the tasks trialled (administered to a similar group of students) before they become exam tasks?*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Yes, always	0	0%
Yes, sometimes	0	0%
Not trialled	20	100%

7.1.2. Staff selection and training for the examination

This part of Survey 1 deals with staff selection for the Examination (materials design and Exam administration) and the available support for materials designers in the form of seminars, workshops and access to resources. The key issues to investigate are criteria for staff selection and availability of training.

The first question in this part, Q12, was asking the respondents if they knew the **criteria for appointment of materials designers and examiners**. This was a multiple choice question with several options to choose from and ‘other’ option to add any other information the respondents felt relevant. Table 7.11 illustrates the responses obtained.

Table 7.11: Responses to Q12: *What criteria are used by the Department in the appointment of materials designers and examiners?*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Teaching experience	4	20%
Experience in examinations	9	45%
Appropriate qualification	12	60%
Competence in English	5	25%
Expertise in Linguistics	15	75%
Reliability ⁴⁷	1	5%
Other	0	0%

⁴⁷ ‘Reliability’ here mostly stands for attitudes (a reliable person would not withdraw themselves from the Examination Board at the last moment, or declare special circumstances). Apart from that ‘reliability’ is synonymous with ‘efficiency’ and ‘experience’ that would allow examiners to function effectively

Table 7.11 demonstrates that there are two criteria the respondents are quite certain about – appropriate qualification and expertise in Linguistics. The third criterion chosen rather frequently is experience in examinations.

The next set of questionnaire items deals with **training for examiners and Exam materials designers**. Table 7.12 presents the responses to Q13 and Q15 about availability of any kind of training for examiners and materials designers.

Table 7.12: Responses to Q13: *Do materials writers get some training in materials design (locally or centrally)?* and Q15: *Is there any training for examiners before the examination?*

		Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Training for materials designers	yes	1	5%
	no	10	50%
	don't know	9	45%
Training for examiners	yes	1	5%
	no	13	65%
	don't know	6	30%

One respondent stated that there was training both for examiners and Exam developers, whereas 9 and 6 people respectively found the question difficult to answer. This may be explained by difference in respondents' understanding of training, or even difference in respondents' expectations of what (if any) training should be provided but all these give reason to conclude that, similarly to appointment of examiners, their training procedures are not made clear enough.

As a possible result of respondents' uncertainty about availability of training for Exam materials developers and examiners, Q14 and Q16 yielded quite predictable responses on the **types of training** provided:

Table 7.13: Responses to Q14 and Q16: *What kind of training is it?*

	Options chosen	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Seminars for materials writers	yes	0	0
	answer missing	5	25%
On-line webinars on materials design	yes	0	0
	answer missing	5	25%
Coordination meeting at the faculty (on materials design)	yes	1	5%
	answer missing	5	25%

How to administer the exam	yes	1	5%
	answer missing	4	20%
How to assess student answers	yes	0	0
	answer missing	4	20%
What level of performance is expected	yes	0	0
	answer missing	4	20%

As the table above demonstrates, almost none of the options presented in Q14 and Q16 were chosen. Only one person stated there were coordination meetings at the Department on issues of Exam materials design, and one person chose the option ‘meetings on how to administer the Exam’. None of the respondents used the ‘other’ option to add another type of training.

7.1.3. Assessment procedures

This part of the questionnaire deals with assessment issues – criteria employed for assessment of oral answers, their weighting and marking procedures.

The analysis and categorization of open-ended responses to Q19 ‘What **criteria** are employed for **assessing** student answers?’ first resulted in a list of criteria (the left column in Appendix 14, p.352-255). Then the data was presented in Table 7.14 and quantified through SPSS to see which criteria were prioritised by the respondents (Figure 7.3)⁴⁸.

Table 7.14: Analysis of open-ended responses to Q19: *What criteria are employed for assessing student answers?*

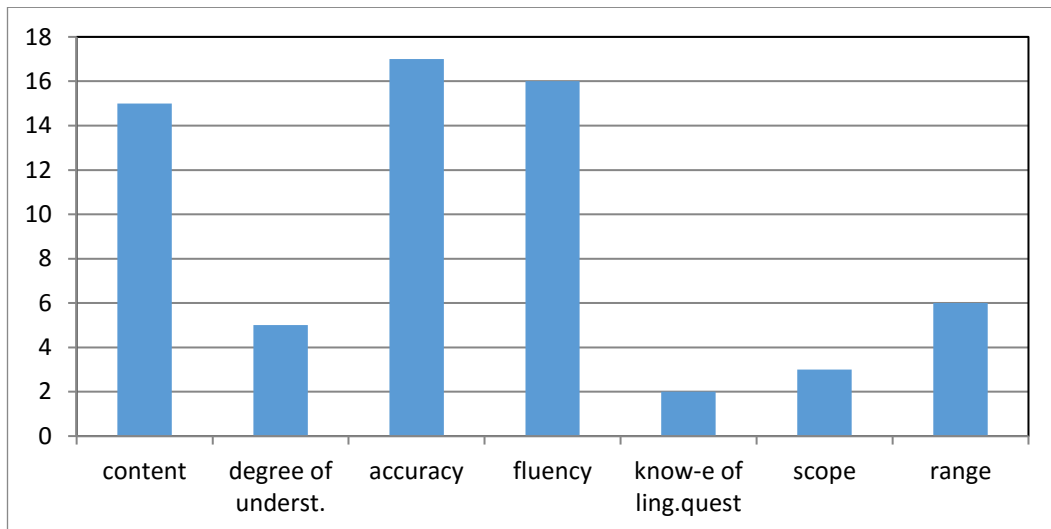
Respondents ⁴⁹	Open-ended responses
<i>Exam_staff_1</i>	Contents, relevance to the topic, accuracy, fluency (depends on the subject)
<i>Exam_staff_2</i>	Accuracy, content, fluency
<i>Exam_staff_3</i>	Knowledge of linguistic questions, degree of understanding of texts, accuracy, fluency, coherence
<i>Exam_staff_4</i>	Accuracy, fluency, scope
Respondents	Open-ended responses
<i>Exam_staff_5</i>	Accuracy, fluency, degree of understanding
<i>Exam_staff_6</i>	In Linguistics: level of knowledge Receptive skills: level of comprehension (% of information)

⁴⁸ The procedure is described in detail in Chapter 6: Research Methodology

⁴⁹ The respondents are Exam designers, examiners and exam administrators – members of staff at the Faculty of FL

	<i>Speech</i> : pronunciation and intonation characteristics, ability to summarize, comment and enlarge on the texts (heard and read), Gr correctness , vocabulary diversity, fluency and spontaneity
<i>Exam_staff_7</i>	Accuracy , fluency , how much was said and what was said
<i>Exam_staff_8</i>	Accuracy , content , variety of language means
<i>Exam_staff_9</i>	Accuracy , fluency , content
<i>Exam_staff_10</i>	Accuracy , fluency , range of vocabulary and structure , content
<i>Exam_staff_11</i>	Accuracy , fluency , content
<i>Exam_staff_12</i>	Fluency , accuracy , content
<i>Exam_staff_13</i>	Content , accuracy (lexical, phonological, grammatical), fluency , range
<i>Exam_staff_14</i>	Content , degree of understanding of the text
<i>Exam_staff_15</i>	Content , linguistic accuracy , tempo
<i>Exam_staff_16</i>	Accuracy , content , fluency
<i>Exam_staff_17</i>	Content , language means, variety of language means
<i>Exam_staff_18</i>	Content , degree of understanding of the text
<i>Exam_staff_19</i>	Accuracy , fluency , content
<i>Exam_staff_20</i>	Accuracy (grammar, lexical), fluency , range , scope

Figure 7.3: Frequency of mentioning the assessment criteria by the respondents



Note: the diagramme presents assessment criteria in the form they were named by the respondents (Table 7.14 above). ‘Degree of understanding’ can be applied to Task 2 and 3 only, because it means degree of understanding of listening and reading texts; whilst ‘knowledge of linguistic question’ can only be applied to Task 1. ‘Scope’ refers to how much Exam taker said, and ‘range’ means range of grammar, vocabulary and phonological means

Figure 7.3 demonstrates that there is some lack of agreement between the respondents about what criteria are employed for assessment. There are some criteria mentioned by nearly everybody: content, accuracy and fluency, which are supposed to be applied

to all three tasks. Range and degree of understanding of a reading/listening text are the second group mentioned, whereas ‘knowledge of linguistic question’ is only mentioned once.

The question about **the weight of each criterion** yielded many ‘don’t know’ responses within each category of respondents. The exception is the most experienced group of examiners (10 years and more) – six people stated that all criteria have equal weight but even within this category difference in opinions is observed – three people were not sure and one person said that the weight of criteria is different.

Table 7.15: Responses to Q20: *Do all criteria have the same weight?*

		Do all criteria have the same weight?			Total
		yes	no	don't know	
Respondents' experience	less than 2 years	0	0	1	1
	2-5 years	0	1	4	5
	6-10 years	1	0	3	4
	more than 10 years	6	1	3	10
Total		7	2	11	20

Then the question was asked about **descriptors** being available for each criterion. Similarly to the previous question, the responses were cross tabulated according to experience of examiners.

Table 7.16: Responses to Q21: *Are there descriptors for each criterion – what is excellent, good, etc.?*

		Are there descriptors for criteria?			Total
		yes	no	don't know	
Respondents' experience	less than 2 years	0	1	0	1
	2-5 years	1	3	1	5
	6-10 years	1	2	1	4
	more than 10 years	1	7	2	10
Total		3	13	4	20

The same tendency can be traced along all experience groups – the majority of respondents chose the ‘no’ option with four people being ‘not sure’.

When evaluating the **usefulness of the existing criteria** for marking student answers and resolving possible disagreement between examiners, 17 out of 19 respondents of all levels of experience chose the ‘don’t know’ option in the questionnaire, with one

person with more than 10 years of experience at the Exam saying that the criteria are helpful:

Table 7.17: Responses to Q23: *Are the existing criteria helpful in resolving disagreement (if any)?*

		Are existing criteria helpful?			Total
		yes	no	don't know	
Respondents' experience	less than 2 years	0	0	1	1
	2-5 years	0	1	4	5
	6-10 years	0	0	4	4
	more than 10 years	1	0	8	9
Total		1	1	17	19

The reasons provided in the open-ended part of the question were that the existing system of assessment criteria is 'too vague'/'too general' and difficult to refer to as 'there is nothing to point at' in case there is some disagreement and examiners want to support their point of view.

The way **the final mark** is given seems to be a logical consequence of the practices described above. The respondents were asked an open-ended question (Q22) about the ways the final mark given to each student. The responses were analysed and categorised to come out with a list of most frequently mentioned ways of reaching an agreement (Table 7.18). It turned out that the only way of grading in the situation under review seems to be discussion of examiners' opinions.

Table 7.18: Responses to Q22: *How do examiners come to agreement about the final mark?*

Respondents	Open-ended responses
Exam_staff_1	Discuss the final mark
Exam_staff_2	Compare their notes and discuss their opinions
Exam_staff_3	Express their opinions on each answer and then discuss the marks
Exam_staff_4	Discuss each answer
Exam_staff_5	Discussion , sometimes voting
Exam_staff_6	By voting
Exam_staff_7	Discuss the answers
Exam_staff_8	Share their opinions
Exam_staff_9	As a result of discussion
Exam_staff_10	Everybody expresses their opinion and then the option everybody agrees with is chosen
Exam_staff_11	Discuss why they think this and not that mark should be given
Exam_staff_12	Discuss what they think about each answer
Exam_staff_13	Discuss what they think
Exam_staff_14	Negotiate it

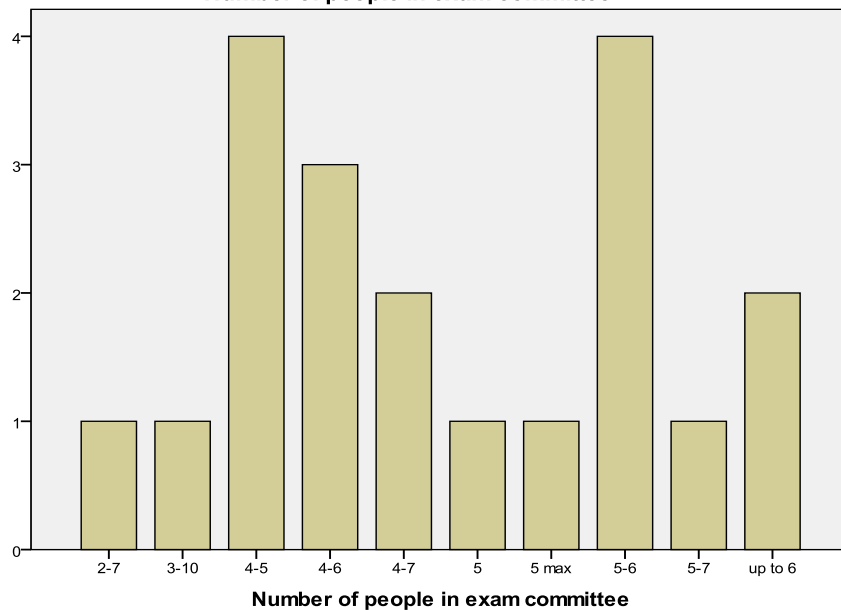
Exam_staff_15	Keep notes and discuss what everybody thinks
Exam_staff_16	Through discussion (after students finish answering)
Exam_staff_17	They keep notes while listening, give a mark; then discuss and come out with the final mark
Exam_staff_18	Discussion at the end of exam
Exam_staff_19	By discussing and sharing their opinions
Exam_staff_20	Discuss each student's answer

7.1.4. Exam administration

This part of the survey deals with various issues of Final Language Exam administration – from the number of people in the Examination Board and number of students taking Exam in one day, to timing issues.

The open-ended responses to Q24 demonstrate that the usual number of examiners is 4-6 people, with the lowest number coming down to two and the highest being ten (Figure 7.4).

Figure 7.4. Responses to Q24: *How many people are there usually present in the actual examining committee?*



Numbers 0-4 on the vertical axe of the diagram show the number of responses provided for each option. As can be seen, the number of Exam Board members most frequently mentioned is 4-6 people.

The **number of students** taking their exam in one day can vary from 8 to 15 with 10-12 being the most common size of the cohort. The number of students taking exam does not influence the number of examiners and only makes the duration of the examination shorter or longer.

The **duration** of the Final Language Examination usually varies between four and seven hours, depending on the number of exam takers and other circumstances (for example, the number of additional questions from Exam Board members), with an average Exam length being 5-6 hours (Table 7.19).

Table 7.19: Responses to Q28: *How long is the examination for examiners (from the very beginning to the very end)?*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
4 hours	1	5%
4-6 hours	1	5%
5 hours	6	30%
5-6	5	25%
5-7	2	10%
6 hours	5	25%
Total	20	100%

Each student spends a certain amount of **time** in the Final Exam room. This time is divided into a) time for preparation and b) time for answering the three questions from Exam card and, possibly, additional questions from Exam Board members.

As can be seen from Table 7.20 on the next page, students can spend from one to three hours in the Final Exam room, with no exact time specified by the respondents or in the documents. The average amount of time is considered to be about 1.5 hours but the real situation for every exam taker depends, as it can be presupposed, on several factors (see part 7.2: Post-Exam Survey).

Table 7.20: Responses to Q29: *How much time (on average) does each student spend in the exam room (preparation time + speaking time)? Please give minimum and maximum time*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
1-2 hours	1	5%
1-2,5 hours	1	5%
1,5 hours	5	25%
1,5-2 hours	3	15%
1,5-2,5 hours	2	10%
1,5-3 hours	2	10%
1hour 20min - 3hours	1	5%
2 hours	5	25%
Total	20	100%

7.2. Post-exam survey for examiners and Exam takers

This survey aims at obtaining additional information on Final Language Exam administration, i.e.:

- actual length of the Examination (including time Exam takers have for preparation and time they spent answering);
- if reference materials are allowed/not allowed; if students and examiners are supposed to keep notes and what happens to the notes after the Examination is finished;
- examiner behaviour and whether it is specified;
- marking process and application of the existing assessment criteria.

Although there were two separate questionnaires – one for examiners (Appendix 8A) and the other for exam takers (Appendix 8B) – the responses to both are presented together, according to the issues covered.

Questionnaire for examiners started with the question about **examiners' degree and experience** in the Examination. All the respondents have doctorate degrees in Linguistics and their experience in the Final Language Exam is two years and above. Exam takers were first asked about the **questions** in their Examination cards and, as described in Chapter 5, all students had different questions on Linguistics and different texts for reading and listening.

The next set of questions in both questionnaires aimed at investigating the **length of the Examination** and whether all students spent the same amount of time in the Exam room. The examiners were asked when the Final Examination started and finished.

Table 7.21a: Responses to the open-ended question ‘Examination started _____’

Time the exam session started	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
9:00	3	75%
9:05	1	25%
Total	4	100%

Table 7.21b: Responses to the open-ended question ‘Examination finished _____’

Time the exam session finished	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
14:40	2	50%
14.30	1	25%
14:25	1	25%
Total	4	100%

Fifteen minutes (Table 7.21b) might be not a big difference but still it is important. It might mean that people treat the phrase ‘exam finished’ differently. Some consider Exam ‘finished’ at the time the marks are announced, others might think it is the time when the last student leaves the Exam room. No directions about how long the Exam session should be were found in the documents. It can be explained by the fact that the number of students taking the Exam may vary (from 9 to 13), so the duration of the Examination cannot be prescribed.

Within the timing issue, the Exam takers were first asked about the **time** they had **for preparation**, i.e. the time between taking the Examination card and starting the oral answer in front of the Examination Board. The responses are presented in Table 7.22a.

Table 7.22a: Exam takers' responses to the open-ended question 'Time you had for preparation'⁵⁰

Time Exam takers had for preparation	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
1 hour	2	18.2%
1 hour (not enough)	1	9.1%
1 hour+	1	9.1%
1,5 hours	3	27.3%
1hour 10 min	1	9.1%
2 hours	1	9.1%
2 hours +	1	9.1%
2,5hrs, was so tired	1	9.1%
Total	11	100%

Another timing issue for the Exam takers to consider was **how long students spent answering** (Table 7.22b). By 'answer' the questionnaire item meant answer to all three questions from the Examination card and answering examiners' additional questions or responding to examiners' comments if there were any.

Table 7.22b: Exam takers' responses to the open-ended question 'Time you spent answering'

Time exam takers spent answering	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
1 hour	2	~18%
45 minutes	4	~36%
30 minutes	4	~36%
15 minutes	1	~9%
Total	11	100%

As can be seen, the differences are: 1-2.5 hours for preparation (Table 7.22a) and ¼ - 1 hour for answering (Table 7.22b). The Final Examination Syllabus states that students are given one hour *for preparation* (Chapter 5), whereas the responses, although quite limited in number, demonstrate that the real situation is quite different. No limitations or any kind of specification have been traced as far as *answering time* is concerned, so different Exam takers had between 15 minutes and 1 hour to answer all 3 Exam questions. It could have happened due to various reasons – examiners

⁵⁰ The figure presents the responses as given by exam takers, so some contain some evaluative judgements

giving clues to some students and not giving clues to others, examiners asking different amount of additional questions/not asking those questions at all; different level of difficulty of Task 1, etc. Whatever reasons were behind this difference in answering time, a 45-minute difference between the minimal and maximum time is rather substantial.

As stated previously, **examiner behaviour** is an issue not quite consistently and clearly presented in the Final Examination Syllabus. It still remains unclear after obtaining the data from Survey 1. No information has been obtained on what examiners are expected/not expected to do before and during the Examination, so the question was further investigated in the survey for examiners and Exam takers. Responding to question about examiner behaviour being specified (Table 7.23) one examiner said it was specified, two said ‘no’ and one chose the ‘not sure’ option.

Table 7.23: Examiners’ responses to the question ‘Is examiner behavior specified?’

	Number of questionnaire respondents who chose this option	Percentage of total number of respondents
yes	1	25%
no	2	50%
not sure	1	25%
Total	4	

Answering the open-ended question about the necessity of specifying examiner behavior, 3 out of 4 examiners responded positively – stating that ‘the conditions should be more or less equal’, ‘not to ask too many questions as a State exam makes students very nervous’, ‘different students get different amount of attention’. One person was ‘not sure’ if it was possible to specify examiner behaviour because it is difficult to predict all situations likely to happen.

An important issue of examiner behaviour at an oral examination is **a possibility to interfere into student answers**. 4 out of 4 responding examiners said they could do it and they did it when necessary. 7 out of 11 exam takers stated there was examiner intervention, whereas 4 said there was none. Both examiners and exam takers were asked about the **reasons of intervention**. The results are presented in Table 7.24 where examiners’ and Exam takers’ responses are compared.

Table 7.24: Examiners' and Exam takers' responses to question 'What were the reasons for intervention?'

Reasons for intervention	Examiners		Exam takers	
	Number of respondents who chose this option	Percentage of total number of respondents	Number of respondents who chose this option	Percentage of total number of respondents
Content covered insufficiently	4	100%	3	27.3%
Language mistakes	1	25%	--	
Other mistakes	2	50%	4	36.4%
Task not achieved	4	100%	1	9.1%
Other	3	75%	10	90.9%

The 'other' reason for intervention, according to examiners, was necessity to ask additional questions because 'not everything the students said was clear', 'questions are to be asked according to the accepted procedure' or 'to check if students can react spontaneously'. As Table 7.24 illustrates, the perceptions of Exam takers are slightly different from those of examiners. Exam takers mostly see intervention as an opportunity 'to know how I know the question' or 'to see if I know other areas'. One student admitted that examiners 'wanted [them] to say what I did not plan to say'.

Table 7.25 deals with different **ways of intervention** and how often they are employed. Similarly to the reasons of intervention, the ways are seen differently by examiners and Exam takers.

Table 7.25: Responses to question 'What kind of intervention was it?'

Kind of intervention	Examiners		Exam takers	
	Number of respondents who chose this option	Percentage of total number of respondents	Number of respondents who chose this option	Percentage of total number of respondents
Correction	1	25%	4	36.4%
Praising/encouraging	--		2	18.2%
Giving a clue	4	100%	3	27.3%
Other kind				
conversation	1	25%	--	
questions	3	75%	3	27.3%
interruption	--		3	27.3%

Neither kind nor amount of intervention is prescribed by the Final Exam Syllabus (Chapter 5), and no clarification has been obtained from the results of the post-Exam survey on if/when intervention is required and what form it should take (clues, questions, direct error correction, etc.). As a possible consequence of very vaguely

described examiner behaviour, the **effects of examiner intervention** are different for different Exam takers – from helping some to disturbing thoughts of others.

Table 7.26: Responses to question ‘*What was the effect/result of intervention?*’

	Number of respondents who chose this option	Percentage of total number of respondents
helped student to improve	3	27.3%
disrupted student’s ideas	4	36.4%
student didn’t understand why intervention took place	2	18.2%
not sure	2	18.2%

As can be seen from the table above, only 3 Exam takers stated that examiner intervention had a positive effect on their answer and helped them to improve. Four students seem to have lost their thought because of intervention, 2 people were not able to understand examiners’ clue or whatever kind of intervention it was.

Another issue which the post-Exam questionnaires aimed at investigating was **marking procedures**. The questions about marking were mostly addressed to examiners with only some items addressed to both groups. First, to get a clear idea of how marking of oral answers takes place, examiners were asked whether they **kept notes** while listening to students. Four out of 4 responding examiners said they did. The data obtained from Exam takers is slightly different. Eight students said the examiners kept notes whereas 3 respondents stated no notes were kept. The difference in responses from the two groups might be explained by difference in understanding ‘note keeping’ – students might have treated this as keeping note of their own answer and examiners might mean the whole cohort of students. The notes are kept for further **negotiation of the marks** between the members of the Examination committee/board: 4 out of 4 examiners said the marks were negotiated before being announced. Final marks caused argument between the examiners with major reasons being content of student answers (3 out of 4 examiners) and language (all 4 examiners).

The existing set of **criteria** is seen by the examiners as being not too **helpful in marking** and resolving disagreement about final marks: 1 examiner did say the criteria were helpful whereas 3 examiners were not sure. It might be partially explained by the absence of descriptors for each criterion: 1 examiner said the criteria were not **clearly written**, 2 were not sure and only 1 respondent said the criteria were clear enough.

The examiners were asked to comment on any aspect of the Final Language Examination and most comments were about assessment criteria and how marking was administered:

Table 7.27: Summary of examiners' comments on the Final Language Examination

Examiner_1	There is a list of criteria which is difficult to apply. I know what accuracy means and what fluency means this is not helpful. The most important seems to be content as they answer a linguistic question. But what if they make language mistakes? Something should be reconsidered about criteria
Examiner_2	I do not think it is possible to assess the answer to the first question and the second question together and announce one mark. It's not clear what we are assessing. Even if it is knowledge – what kind of knowledge is it? The criteria can only be applied to question 2 and even there it is not clear how to apply them. I am not happy with this
Examiner_3	I am not happy with the new format – there is only one task when they answer a linguistic question. It is not clear at all how the criteria apply to this task. What shall I assess? Content or accuracy or fluency or everything? No, it's not clear
Examiner_4	I believe that our State exam doesn't really check students' language skills. It rather checks their ability to memorise a lot of theoretical material and their experience as school teachers

As described in Chapter 6: Research Methodology, open-ended responses were analysed and categorized (p.135). Table 7.27 illustrates that there are three most common responses: the criteria are difficult to apply and therefore not very helpful; assessment focus is not very clearly defined. In some way similarly to the examiners, students, when asked about **helpfulness of the criteria for Exam preparation**, said they were not (10 out of 11 respondents) and only one student said the criteria were helpful.

The data obtained through the post-exam survey is in good keeping with the data from Survey 1, namely the information about the Final Exam administration procedures. The results presented above demonstrate that:

- as **timing** is not clearly defined and described in the Final Examination Syllabus or any other document. It is seen as rather flexible which results in different amount of time for preparation (from 1 to 2.5 hours), different time students spend answering (from 15 to 60 minutes) and, in the end, different times they spend in the Examination room;

- the data from the post-exam survey demonstrate that **examiner behaviour** is not specified, which results in different amount of examiner intervention (from none to unlimited number of corrections and additional questions);
- **assessment criteria** mostly come from examiners' experience and expertise rather than are defined by any official document. The criteria are not seen as helpful both by the examiners and Exam-takers.

Chapter 7 contains data obtained from various stakeholders – Exam designers, examiners, administrators and Exam takers. The chapter aims to present and classify major findings on Exam design and administration: from selection of Exam tasks to announcing final marks. This task was achieved by presenting the results of 2 surveys – Survey 1 for Faculty of FL staff who are major decision makers on all Exam dimensions, from planning to implementation. The data obtained through Survey 2 aims to look at the Examination from a slightly different perspective, i.e. to describe what exactly goes on in the Exam room as seen by examiners and Exam takers. Two other important dimensions of the Examination – its content and format – are presented and discussed in Chapters 9 and 10.

Chapter 8

Discussion of findings on Research Question 1: procedures of Exam design, piloting and administration as seen by different stakeholders

This chapter provides discussion of the qualitative and quantitative data obtained through Survey 1 for Exam designers and examiners and Survey 2 for examiners and Exam takers. The collected data is discussed alongside the following lines:

- Exam materials design, including issues of content selection, defining Exam format, Exam designer training, piloting of Exam materials;
- Exam administration – assessment criteria and their use, marking student answers.

Then Chapter 8 discusses strong and weak points of the design and administration of the Final Language Examination and the ways they contribute to Exam validity, reliability, authenticity and practicality - the key parameters of test evaluation considered in the literature review (see Exam evaluation checklist: Chapter 4, pp.107-109). This allows to single out some threats to the current Examination that can be minimized in the process of designing alternative materials for the Final Language Examination.

8.1. Final Language Examination materials design

The data obtained from Survey 1 for Exam designers and examiners (Chapter 7; Appendix 7) supports the information that can be found in the key Exam documents: Final Examination Syllabus and Dean's orders (Chapter 5). The Examination is developed internally by the Faculty of Foreign Languages staff, with no other stakeholders like school teachers of English or staff from other universities involved.

This research did not aim to find any empirical evidence on whether involvement of school teachers or representatives of school administration in the process of Exam task design would contribute to validity and reliability of those tasks. Nevertheless, it can be presumed that feedback from school teachers of English on Exam task difficulty and appropriacy to teachers' job might be relevant. Some feedback from school English teachers obtained in

this research about the current Final Exam content and format demonstrated that school teachers have their own vision of the Exam and some ideas of possible changes that can be quite useful in the design of the Examination (Chapter 9).

The Exam design procedures are not described in detail in the Exam documentation (Chapter 5) and remain vague after analysis of the obtained data. The respondents testified that Exam design does not follow any clear guidelines. On the one hand, employing task and text types that ‘proved to be useful’ and those that ‘were chosen many years ago’ (Table 7.5; Figure 7.1; Tables 7.6, 7.7) each year makes designers’ task easier and, probably, does not require detailed explanation of each stage of materials design. On the other hand, absence of a clearly defined procedure of Exam design might be suggestive of some lack of planning in this area and even some lack of understanding of what this stage might be like. The data obtained on Exam task moderation might contribute to the concern about some lack of understanding of Exam design procedures. In this research, moderation, after Alderson (1995), is viewed as:

Assembling [text] items into a draft test paper, for the consideration of the formal committee. The task of this committee is to consider each item and the test as a whole for the degree of match with test specifications, likely level of difficulty, possible unforeseen problems... . The committee do not simply read the test and its items: they must take items as if they were students (1995: 62-63).

Responses provided to the question about Exam task moderation (Chapter 7: Tables 7.8, 7.9) demonstrate some lack of agreement of what Exam design must include, whether task moderation is essential and even what moderation is. The situation with piloting Exam materials seems to be much more straightforward (Chapter 7, Table 7.10): 20 out of 20 respondents said Examination materials were not piloted. It might mean that potential problems are ‘not ironed out before the major trial’ (Alderson, 1995: 73); task types are not calibrated according to their level of difficulty and no corrections can be introduced in description of student expected performance.

In the context under study, piloting may be quite problematic due to absence of piloting population: all final year students are supposed to take the Final Examination, so they cannot be given sample tasks for piloting. It might be problematic to find a suitable cohort of students who would pilot the Exam tasks in other universities due to technical reasons, difference in programmes of studies and, mainly, due to the fact that piloting is not considered as essential step in materials design and therefore is not treated as

obligatory by university staff. At the same time, moderation of Exam tasks is easier to perform, that is why a clear definition of Exam materials design might be an initial step in introducing moderation at the Faculty of Foreign Languages.

Another factor that contributes to the opinion of Exam materials design being a vague issue is training for examiners and Exam designers. As Tables 7.12, 7.13 (p.164) demonstrate, the examiners and Exam designers participating in this research chose different options on whether training takes place, with half of respondents stating that training is available for task designers, but not for examiners, and half of respondents being not sure at all.

Table 8.1. summarises strong and weak points of Final Language Exam design as seen by various stakeholders.

Table 8.1: Examination design at university

Advantages	Disadvantages
<ul style="list-style-type: none"> • a team of experienced Exam designers • materials design procedure well-established • materials design process is not time consuming 	<ul style="list-style-type: none"> • no clearly defined criteria for task and text selection • no moderation or piloting of Exam tasks • no training for examiners, Exam designers, assessors

The major advantage of this stage, as can be seen from the obtained empirical data, is a well-established team of task designers with significant experience (Chapter 7: Table 7.25e). Such a team knows the requirements to task difficulty, choice of tasks and texts and knows quite well what can be expected of Final Exam takers in terms of their performance.

From another perspective, some of the advantages above can be considered as weak points of Exam design. Thus, knowing what task types work well allows for the same choices year after year. Having quite a limited awareness of possible alternative task types limits a possibility of new task types being sought and employed. Absence of clearly defined criteria for choosing task and text types might lead to convenience choices. As there are no requirements in the State Standards on exam materials design in general and choice of exam tasks in particular, necessity to introduce changes in the Exam content, format or administration might be not straightforward at all. Exam design procedures that are vaguely defined in the Final Examination Syllabus might be a threat to Exam reliability, according to H.D.Brown (2004). Absence of

moderation and piloting may have the same effect, according to Alderson (1995) because exam developers have no opportunity to see

‘the degree of match with the test specifications, likely level of difficulty, possible unforeseen problems, the overall balance of the test in paper’ (1995: 63).

Another issue threatening the Final Exam reliability is absence of training or other co-ordinating events for Exam designers, examiners and Exam administrators. Under such circumstances, Exam designers lack an opportunity to vary and enlarge the range of tasks and texts by, for example, learning of experience in language testing in general and language testing for teachers. Absence of any co-ordination meetings for assessors might, in some way, result in lower inter-rater reliability (Chapter 4) due to possible differences in assessors’ perceptions and expectations of student performance.

As can be seen, lack of clearly defined Exam design procedures is considered in this study mostly as a threat to Exam reliability. Although the list of threats is rather substantial, standardisation of Exam materials development is seen in this study as a way of improving the current situation.

8.2. Administration of the Final Language Examination

Final Language Exam administration is the issue that was mostly investigated through obtaining empirical data because the Final Exam Syllabus gives quite vague guidelines on Exam administration. Thus, data from Survey 1 for examiners, Exam designers and administrators, and post-exam Survey 2 for examiners and Exam takers were analysed to get as clear a picture as possible of:

- timing issues: time for preparation, time students have for answering;
- examiner intervention and examiner behaviour;
- use of resources at the Exam;
- assessment criteria and marking procedures.

Empirical data obtained from examiners and Exam designers in Survey 1 demonstrated that *timing* was an issue not thoroughly described in the Exam Syllabus or agreed between examiners and Exam administrators. As presented in Chapter 7, Exam takers have different amount of time to prepare their answer in the Exam room – from 60min

to up to 2hrs; spend different time answering questions from their Exam cards. Post-exam Survey 2 for examiners and Exam takers cast more light on the issue. Approximately 40% of Exam takers were preparing their answer for 1hr or a bit longer (1hr 10min), whereas 30% of students had 1.5hrs at their disposal, and another 30% of students had 2hrs and more to plan and prepare their answers (Chapter 7: Table 7.22a). The amount of time Exam takers spent answering in front of the Exam Board was also significantly different, according to Survey 2. It varied from 15min to 1hr (Chapter 7: Table 7.22b). This research did not aim to investigate whether different amount of time that students got for preparation and answering influenced their performance and final marks at the Exam. However, theoretical and empirical insight into timing issues in language testing (e.g. Kane, 2010; Xi, 2010) suggested that such unequal conditions can be considered a threat to exam reliability and reliability of final marks (Alderson, 1995; J.D.Brown, 2000; Hughes, 2003; H.D.Brown, 2004).

Timing issues might be in some way caused by unspecified *examiner behaviour*. As stated in Chapter 5, no indication was found in the Exam syllabus on any rules and regulations for examiners. With the current Exam format (discussed further in Chapter 10), standardization of examiner behaviour seems quite problematic. The responses to Survey 1 and 2 demonstrated that examiners usually act in the way they find necessary, and very often examiner behaviour depends on what an examiner thinks is right. This results in different examiners acting in different ways, i.e.:

- giving different number/types of clues to Exam takers, or giving none;
- making/not making corrections in what students are saying at the Exam;
- asking/not asking additional questions in and outside the Exam task area;
- interrupting/not interrupting student answer for any reason (Chapter 7: Table 7.25).

As a result of non-standardised examiner behaviour, according to post-exam Survey 2, different Exam takers got different amount of intervention from examiners: from direct correction to interruption (Chapter 7: Table 7.25). It is quite interesting that examiners and Exam takers saw reasons for intervention differently: whereas examiners said they had interfered with students' answers mostly because of mistakes they were making, Exam takers very often did not see this as a reason (Chapter 7: Table 7.24). The effects of intervention were also seen quite differently by both parties: only 3 out of 11 students admitted that examiner intervention had helped them to

improve, whereas 4 said it had disrupted their ideas and other 4 did not realise why intervention was taking place (Table 7.26).

In some way, such a situation is seen as predictable. In case of subjective oral assessment with integrative tasks employed, the possibility of predicting all possible situations that can happen at the Exam is rather low. It makes specifying examiner behaviour quite a challenging task. At the same time, as publications in language testing claim (Alderson, 1995; J.D.Brown, 2000; Hughes, 2003; H.D.Brown, 2004), and some national and international examination bodies put into practice, some standardization is possible through:

1. task pre-testing, that helps to:
 - make sure if tasks ‘elicit the intended sample of language, whether the marking system is usable and whether the examiners are able to mark consistently’ (Alderson, 1995: 76; Heaton, 1995; McNamara, 1997)
 - elicit different situations and in this way to predict possible examiner behaviour;
 - limit the ‘freedom’ (Hughes, 2003) of test takers by specifying what exactly is expected of them;
2. thorough description of student expected performance;
3. choice of task types that regulate examiner behaviour (e.g. an oral interview with pre-designed interview framework and/or visual input; a set of questions to ask on the article that Exam takers read; a pre-designed statement to provoke a discussion)
4. providing descriptors for assessment criteria.

Assessment criteria for the current Final Language Examination is another area that requires more thorough planning and description. Similarly to other Exam administration issues, the assessment (scoring) system gets very little attention in the Exam syllabus. It was hoped that Survey 1 and 2 would cast more light on what criteria are employed for assessment and how marking system works. In Survey 1, the examiners and Exam designers were first asked what criteria were employed for assessment (Chapter 7: Table 7.14). Accuracy, fluency and content were those most frequently referred to with others like ‘degree of understanding [of listening and reading texts]’, ‘knowledge of linguistic questions’, ‘scope and range [of lexical and grammatical means]’ being mentioned fragmentarily (Chapter 7: Figure 7.3). None of the respondents distinguished between Task 1 (Linguistic question) and Task 2-3

(Listening and Reading) in terms of assessment criteria applied to assess student performance. Judging by the samples of Exam questions (Chapter 5, p.115), Task 1 and Tasks 2-3 differ significantly in input, assessment focus, expected performance and, therefore, are expected to differ in assessment criteria. Nevertheless, no information could be obtained in how assessment system works and whether all criteria have the same weight in all three tasks (Table 7.15). Input that is different for each student, together with different amount of time Exam takers have for preparation and oral answers can only add to absence of clearly defined criteria and result in lack of consistency in examiners’/ marker’s work.

Table 8.2 summarises the strong and weak points of Exam administration as seen through empirical data from Surveys 1 and 2.

Table 8.2: Administration of the Final Language Examination at university

Advantages	Disadvantages
<ul style="list-style-type: none"> • continuity between progress and Final assessment • Exam administration procedure is familiar to Exam takers 	<ul style="list-style-type: none"> • no description of administration procedure • timing issues are not specified; timing is not observed • examiner behaviour is not specified • vague assessment criteria with no descriptors; no clarity about weight of each criterion • the same set of criteria is applied to 2 different task types

The obtained data on Exam administration demonstrates that the current Examination suffers from relatively poor inter-rater and intra-rater reliability that is considered a serious issue for all language tests (e.g. Alderson, 1995; Lumley, 2002; H.D.Brown, 2004; Ling, 2014; Kuiken&Vedder, 2014). In the situation under study, several factors may contribute to low rater reliability, with the major ones being, probably, the format and content of the Final Language Examination. The task types employed at the Exam presuppose extended open-ended responses to input that is different for each Exam taker. Input for each of the 3 tasks cannot be considered equal that makes expected performance different, too. Thus, linguistic questions in Task 1 presuppose different output; Tasks 2 and 3 provide texts for reading and listening that are not at the same level of language and conceptual difficulty and, as a result, cannot always stimulate oral production in equal ways. In addition to this, the existing list of assessment criteria does not contribute to reliability of Exam results: there are no descriptors for each

criterion and no distinction between assessment criteria for the purely knowledge-oriented Task 1 and Tasks 2-3 that aim to assess Exam takers' communicative ability.

Apart from absence of a transparent system of assessment criteria that might threaten rater reliability, there is another factor that might contribute to this issue. Absence of rater and examiner training that is considered crucial by many (e.g. Alderson, 1995; Lumley, 2002; H.D.Brown, 2004; Xi, 2010; Ling, 2014) seems to deepen the problem of quite low rater reliability for the current Final Language Examination.

Standardisation can be considered a complicated issue for oral examinations, where the format presupposes long open-ended responses, whether it is standardisation of input, standardisation of expected performance, standardisation of conditions for exam takers, or standardisation of examiners'/ markers' work. Changes in only one of the above dimensions are highly unlikely because they immediately involve changes in others. Thus, standardisation of expected performance is problematic without standardising input for each task; standardisation of conditions for Exam-takers are unlikely without changing the format of the Exam under study; standardisation of examiners' and markers' work seems impossible without development of a transparent system of assessment criteria for each Exam task.

Apart from considerable threats to the current Exam reliability, the administration issues discussed above influence Exam practicality. As Final Exam Syllabus (Chapter 5) and empirical data from Survey 1 demonstrate (Chapter 7: Table 7.19; Figure 7.4), an Examination for on average 10 students takes up to 6 hours and involves on average 5 people. The roles of the people involved are not defined and there is no division of responsibilities, so all 5 (and sometimes 6) examiners first listen to all students and then mark and discuss their answers. Due to the oral form of the Exam and its format, timing issues – time for preparation and oral answer for each student – seem quite difficult to be defined precisely. All these, together with the vaguely defined assessment and marking procedure make the current Final Language Examination not quite practical (e.g. Heaton, 1995; Hughes, 2003; H.D.Brown, 2004).

The findings on different dimensions of the current Final Language Examination were expected on the one hand, and not quite expected on the other. The obtained data on the current Final Language Examination provided empirical evidence to the Exam

description based on documents – its design, content and format, and administration and demonstrated that a lot of issues remain undefined. At the same time, responses to some questions (for example, questions about Exam specifications, moderation and piloting of the Exam tasks, or questions about assessment criteria in Survey 1) revealed even deeper misunderstanding of Exam design issues. The responses of Exam takers (Survey 2b) appeared to be more negative than expected – almost all 11 respondents felt negative about various dimensions of the Exam, mostly about its administration.

The data obtained on the current Final Language Examination demonstrated some disagreement between what is going on in the studied context and contemporary practices of language assessment of language teachers. On the one hand, the issues revealed through the empirical data are in good keeping with concerns about language examinations that were described by quite a number of authors. These issues manifest through threats to Exam validity, reliability, authenticity and practicality. On the other hand, it can be stated that the approach to assessment of language competence of English teachers (university graduates) under study does not fully fit the existing requirements to language testing, i.e. test materials design, selection of content and format, administration and assessment and marking procedures (e.g. Alderson, 1995; Hughes, 2003; Lumley, 2002; Brown, 2004; Kuiken & Vedder, 2014; Xi, 2014). The present empirical data and analysis of documents on the current Final Language Examination reveal vulnerability of various dimensions of the Exam – from its design to marking of student answers.

Chapter 8 discussed two key dimensions of the current Final Language Examination – its design and administration. The major purpose of this was to investigate the strong points of the Examination, and also possible threats to its validity, reliability, authenticity and practicality that may be caused by the current Exam practices.

Chapter 9 presents data on another important aspect of the Examination: its content and format. The data from different stakeholders is discussed in Chapter 10, aiming to see how current Exam content and format contribute to its validity and authenticity.

Chapter 9

Findings on Research Question 2: How relevant is the Exam content to the language needs of practicing English teachers? What are the language needs of language teachers in Russia?

This chapter is based on three sets of data. The chapter starts with presenting the findings from Survey 1, for the Exam designers and examiners, on the content and format of the current Final Language Examination. Then the chapter introduces the qualitative data obtained through teacher interviews. The interviews dialogued with school teachers with different backgrounds and teaching experience, focusing on their opinions on the content and format of the current Final Language Examination and its relevance to their job.

Then the chapter presents the results of language Needs Analysis of English language teachers in the Tula region. The data obtained through 2 surveys for practising and trainee teachers of English aimed to cast light on the language skills and activities that teachers are involved in regularly in their everyday professional practices.

The visions of the three groups of stakeholders: Exam designers and examiners, Exam takers, and school teachers are then compared and discussed in Chapter 10.

9.1. Final examination content and format

As stated in Chapter 5, no detailed information has been found in the reviewed documents about the Final Language Exam focus, i.e. skills and sub-skills under assessment, level of expected performance of Exam takers. So, the examiners and Exam designers were asked a question about the **focus of assessment** (Q18). Quite predictably, the majority of Exam developers see *linguistic competence* and *receptive skills in general English* as key areas of assessment at the Final Examination: 20 out of 20 respondents either ‘strongly agree’ (11 people) or ‘agree’ (9 people) that graduates’ linguistic competence is assessed; 20 out of 20 respondents with a bit different balance (3 people strongly agreeing and 17 agreeing) say that listening and reading are in the focus of assessment. The situation is a bit more complicated with

productive skills: whereas 8 people ‘agree’ or ‘strongly agree’ that productive skills are tested at the Exam, there are 12 people who preferred the ‘neutral’ option.

Table 9.1: Responses to Q18: *In your view, what is the focus of assessment?*

	Options chosen	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Linguistic competence	strongly agree	11	55%
	agree	9	45%
Receptive skills (general English)	strongly agree	3	15%
	agree	17	85%
Productive skills (general English)	strongly agree	1	5%
	agree	7	35%
	neutral	12	60%

A considerable amount of neutral responses for productive skills being in assessment focus might be explained by the format of the Exam. Out of two productive skills – speaking and writing – only speaking is assessed. In all three Examination tasks students are expected to speak, and all three tasks involve prepared monologue with no or very limited element of spontaneity. In Task 1, a linguistic question, only content is assessed with no emphasis on accuracy, fluency, range of language means, etc. which becomes clear from the assessment criteria employed (Chapter 7: Table 7.14). The other two Examination tasks involve the same skills – retelling and summarizing what was heard/read. The situation with *writing skills* is even more contradictory – although students are supposed to prepare all their answers in writing before speaking in front of the examination committee (see Chapter 5), the notes are not marked and in any other way taken into consideration when the final mark is given.

Receptive and productive skills in professional English produced even more varied responses (Table 9.2). While some respondents (2 and 1 respectively) agreed that these skills are assessed at the Examination under study, some people (9 and 3) were neutral which might demonstrate their doubt. 8 respondents out of 20 did not think that receptive skills in professional English were in the focus of assessment, and 15 people disagreed that productive skills in professional English were assessed.

Table 9.2: Responses to Q18: *In your view, what is the focus of assessment?*

	Options chosen	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Receptive skills (professional/teacher English)	strongly agree	0	0
	agree	2	10%
	neutral	9	45%
	disagree	6	30%
	strongly disagree	2	10%
	answer missing	1	5%
Productive skills (professional/teacher English)	strongly agree	1	5%
	agree	0	0
	neutral	3	15%
	disagree	9	45%
	strongly disagree	6	30%
	answer missing	1	5%

Although the difference in opinions between people talking about the same examination might seem surprising, an explanation might be found in the absence of a directly stated Final Exam purpose and focus. Some texts for listening and reading do deal with issues of education and upbringing⁵¹, but at the same time, there is no clear description of how topics should be represented and balanced at the Exam and whether assessment focuses on general English, professional English, or both.

Some similarity of opinions is observed in the respondents' view of *vocabulary and grammar* being in the focus of assessment at the Final Language Examination (Table 9.3). According to the responses, there is definitely no focus on professional vocabulary in the Examination. The vast majority of respondents (17 people out of 20) do not see Classroom English as an assessment area at the Examination although three respondents chose the 'neutral' option.

Table 9.3: Responses to Q18: *In your view, what is the focus of assessment?*

	Options chosen	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
General vocabulary	strongly agree	2	10%
	agree	18	90%
	neutral	0	
	disagree	0	

⁵¹ With other texts being about ecology, technology, financial issues and none of them dealing with learning and teaching foreign languages

	strongly disagree	0	
	answer missing	0	
Professional/teacher vocabulary	strongly agree	0	
	agree	0	
	neutral	3	15%
	disagree	14	70%
	strongly disagree	3	15%
	answer missing	0	
Classroom English	strongly agree	0	
	agree	0	
	neutral	3	15%
	disagree	11	55%
	strongly disagree	6	30%
	answer missing	0	

9.2. Exam content as seen through interviews of school teachers of English

As presented in Chapter 6: Research Methodology, the interview followed a specially designed interview framework and was conducted in groups with only one teacher having expressed her wish to be interviewed individually. The purpose of the interview was to get opinions of professionals who are not directly involved in the Final Language Examination – neither in its design, nor marking or any other procedures. At the same time, all interviewees passed the Final Exam in a similar format and now, having different amounts of work experience, can think of applicability of the Exam tasks to their job as English teachers at school.

The interview framework (Appendix 11) was designed so that the interviewees could first talk about the routines they are involved in every day as teachers of English and then discuss the Final Language Examination tasks: their difficulty level, appropriacy and applicability to teacher job. Therefore, the framework included:

- the introductory part with a list of everyday activities for the teachers to choose
- a set of questions about Linguistic, Reading and Listening parts of the Examination
- two questions about teachers' vision of the Examination – what parts it should consist of and which parts (if any) can be excluded or reshaped in the current Final Examination.

The introductory talks in all four groups when the interviewees chose what they do at least once a week from the list of activities demonstrated that the most popular ones,

quite predictably, are conducting classes; preparing for lessons, including selection of teaching materials; reading on ELT issues.

Table 9.4: Teachers' responses to the question *'What activities are you involved in at least once a week?'*

Activities	Number of respondents who chose this option	Percentage of total number of respondents
Preparing for lessons	9	81.8%
Browsing Internet for language activities	8	72.7%
Reading books on Linguistics	1	9.1%
Reading books on ELT	7	63.6%
Write about linguistic issues	0	-
Write about ELT issues	0	-
Conduct classes	8	72.7%
Discuss teaching issues with colleagues	4	36.3%
Give feedback to students	8	72.7%
Read reference materials	5	45.5%
Make presentations	4	36.4%

Reflection on teacher everyday professional activities involved the interviewees into reflection on the current Final Examination tasks. The responses about the Linguistic part of the Examination were first analysed and divided into sections (Figure 9.1).

Figure 9.1: Teachers' opinions on Exam Task 1: Linguistic knowledge

Interview 1	M:	The linguistic part is appropriate - I think. Of course, the teacher must know the subject she teaches and that's why they [linguistic questions from Task 1] are absolutely appropriate
	A:	Because we can be asked questions from our students concerning why... this or that phenomenon is used in the language and it is important to be able to answer...so these questions are appropriate
	AN:	I don't know if it's appropriate. I'm not sure teacher can use so deep knowledge working at school. Of course, teachers should know a lot but I think such tasks are more for university teachers (...) for people who want to become linguists. It's too deep for school, pupils do not need it.
	O:	I thought about it when I was preparing for the Exam. I thought ...oh, it's too difficult to explain. You know this and you know that but when you explain the difference between Present Simple and Present Continuous and you see that they [learners] do not understand. And what you know [points at everybody] does not help.
Interview 2	N:	Well...no doubt we need it. From the linguistic point of view. But from the point of view of teaching...it's not too good. (...) I do not remember time I needed it. It's a bit too deep, too detailed. We do need it at a simpler level.
Interview 3	M:	I think a teacher must have a certain background knowledge of language theory. It's [exam questions] probably too much...it's more than needed at comprehensive school
	T:	I think this knowledge is important. We must know these things. But I do not remember I ever used this knowledge. (...) I have to compare with Russian when I teach Grammar. And I explain the difference. But I never go very deep... they do not need it.

When the respondents' opinions on the linguistic part of the Examination were analysed and compared, major strengths and weaknesses of this part were singled out, as seen by the respondent teachers:

Table 9.5: Interviewees' opinion on the linguistic part of the Final Language Examination

Assessment of the Linguistic part of the Final Language Examination	
Support	Criticism
Interview 1	
<p>The teacher must know the subject she teaches and that's why they [materials] are absolutely appropriate (M1, 9 years of experience). ** Teachers should know a lot... (AN) ** ... teachers must know theory (O, 7 years of experience)</p>	<p>I am not sure that teachers can use so deep knowledge working at school.(...) This is too deep for school, pupils do not need it. (AN, 5 years of experience) ** ...And what you [teachers] know about [Grammar] does not help [in teaching Grammar to learners]. ...it [theory] should be less complicated... or more practical (O)</p>
Interview 2 (individual)	
<p>No doubt we need it. From the linguistic point of view (N, 19 years of experience)</p>	<p>From the point of view of teaching practice... it's not too good. It is a bit too deep, too detailed. We do need it at a simpler level. I do not remember the time when I needed it (N)</p>
Interview 3	
	<p>It's (meaning the exam questions) probably too much ... it's more than needed at a comprehensive school. (M2, 7 years of experience)</p>
Interview 4	
	<p>I feel they are a bit too complex... (M3, 5 years of experience) ** We need it but not so... don't know... not in such detail. If we are going to work at school. I would exclude them from exam. We have Theoretical Phonetics, Theoretical Grammar, Lexicology... what else... in the end of each we have an exam. It's enough! (AL, 2 years of experience)</p>

The responses in Table 9.5 were colour-coded to highlight the dimensions which were touched upon by the interviewees. As can be seen, positive commentaries mostly deal with the importance of the linguistic knowledge to a language teacher (highlighted in green) and, therefore, necessity to master this knowledge. Criticism is mostly caused by an unnecessarily deep insight into linguistic theory that is expected of the graduates (highlighted in blue). Related to it is the interviewees' concern about applicability of such detailed theoretic knowledge at a secondary school level (highlighted in light

blue). Last but not the least weakness of the Examination tasks from the teachers' point of view is the focus on the linguistic knowledge but not on an ability to teach this knowledge at school⁵² (highlighted in yellow).

The same procedures were applied to the interviewees' responses about the **Reading** part of the Final Language Examination. The responses were first analyzed and then classified into positive and critical. The responses were colour-coded to summarise the major issues that the teachers touched upon. Below are the samples of opinions from Interviews 1-4 where the respondents (teachers of English) reflect on the Reading and Speaking task of the Final Language Examination (Figures 9.2-9.4).

Figure 9.2: Teachers' opinions on Exam task appropriacy (Reading/Listening and Speaking)

Interview 1 A; O: We conduct classes but we never discuss them... or something else in English.
 O: We have nobody to discuss it with. We... our colleagues who are ... older...
 INT: Colleagues with more teaching experience?
 O: Yes! They cannot discuss these issues in English as they never had Methods⁵³ in English
 M1: I would say we discuss things in Russian but use a lot of English terminology

Interview 3 INT: (...) Do you often do such tasks?
 M2: I never retell articles in English because there is nobody to listen (smiles). I retell things in Russian.
 T: Never retell things... The task [points at the task sample] may be good but we do not need it

Figure 9.3: Teachers' opinions on the choice of Exam texts

Interview 1 M1: I think texts are good. They are quite difficult...quite challenging
 AN: I agree, the level of English is OK, but...
 O: They have nothing to do with teaching. Sorry, I interrupted...
 AN: No problem. I wanted to say they are quite different, you know...if I got this text [points at the sample] about carbon dioxide I don't know what I would say
 INT: Do you mean that...
 AN: They should be closer to our background
 O: Can I say? I think I understand AN. The texts should be closer to our profession

Interview 2 N: I would change the topic. It is so...[points at the sample]

⁵² The concept described in literature as Teacher Language Awareness (see Chapter 3)

⁵³ O. refers to her TESOL course at college and university

INT: This topic is for this text only. This is just a sample. Texts are all different, so topics are all different too.
 N: Are they? Well, if the topics are good, this [the task] is absolutely possible. This text is not too good. I could hardly read it.

Interview 4 M4: I want to say about the texts
 INT: The texts you deal with?
 M: Those for the exam. I think articles chosen for an exam discussion must be taken from authentic up-to-date sources...
 INT: [pointing at the sample text] This is an article from The Guardian, by the way
 M4: Really? Sometimes the texts we read are really old-fashioned
 AL: I would say – the task is all right but texts can be different...from newspapers, magazines. It could be fiction

Figure 9.4: Teachers’ opinions on speaking skills assessed at the Exam

Interview 4 AN: Speaking tasks check ability to understand the message of the text. To my mind, future teachers should be able to speak on a lot of topics
 O: Speaking is not only retelling
 INT: Oh, I agree with you
 O: I like more what we did at college. It was speaking.

The major strengths and weaknesses of the Reading and Speaking task are summarized in Table 9.6.

Table 9.6: Interviewees’ opinion on the Reading and Speaking part of the Final Language Examination

Support	Criticism
Interview 1	
<p>I think the texts are good. They are quite difficult... quite challenging... ... it’s quite good to be able to retell what you read so other people understand... (M1) **</p> <p>It is appropriate... because it evaluates our ability to understand what we read and also highlight the main ideas of it, the main point of it... to say what you think of it... (M1) **</p> <p>...future teachers should be able to speak on lots of topics (O)</p>	<p>Speaking is not only retelling. I like much more what we did at college. It was speaking (O) **</p> <p>They [texts] have nothing to do with teaching. The texts should be closer to our profession (O)</p>
Interview 3	
<p>It’s important to be able to retell what you read so that other people understand it. (T, 9 years of experience)</p>	<p>Never retell things. ... The task [Exam task] may be good but you do not need it (T) **</p> <p>I never retell articles in English because there is nobody to listen (smiles). I retell things in Russian (M2)</p>
Interview 4	

It's essential to know how to sum up the idea and single out the most important facts. As well as giving your own opinion, being able to support it with the appropriate arguments (E)	Sometimes texts we read are really old-fashioned. (M4) ** ... texts can be different. The texts can be from newspapers, magazines. It can be fiction. It's not absolutely necessary but it can be done (AL)
--	---

As Table 9.6 demonstrates, the major strength highlighted by the respondents is an appropriate level of difficulty for reading texts (highlighted in yellow) and the importance of retelling skills for language users in general and English teachers in particular (highlighted in green). The criticism expressed mostly deals with limited skills under assessment (highlighted in purple), non-sufficient range of text types or out of date publications employed (highlighted in grey) and the tasks being irrelevant to the English teacher job (highlighted in light blue). One interviewee's concern dealt with the time the texts were published, as some of sample texts she read seemed out of date.

The parts of all four interviews where teachers discussed the Listening and Speaking part of the Final Language Examination were analysed in the same way as the Reading and Speaking part. The responses were classified into positive and critical.

Table 9.7: Interviewees' opinion on the Listening and Speaking part of the Final Language Examination

Support	Criticism
I would leave everything as it is (M1) ** ...the topics were chosen in a right way... (A)	Some professional component can be added. (A)
It is rather appropriate (M2) I think the task is good. (M3)	

Table 9.7 demonstrates that the interview questions about the Listening and Speaking part of the Examination did not provoke a big discussion. The respondents who expressed their opinions on Listening tasks considered these tasks as good and the topics as relevant. One of the respondents emphasized the importance of a professional component in tasks but did not specify how exactly she saw it: 'I know it is difficult... but maybe to show some part of the lesson... or to describe what you are doing and why' (Interview 1) which seems not to refer directly to the Listening section. In general, Listening and Speaking part is seen by the respondents as quite appropriate to the teacher job.

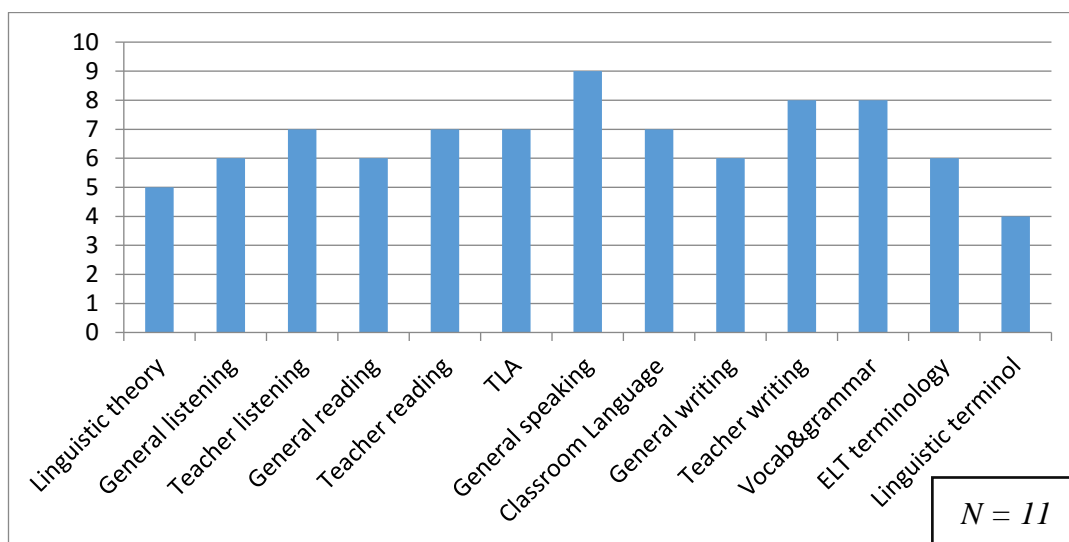
The last issue under discussion in Teacher interviews was a possible format of the Final Language Examination as teachers see it. When asked about possible changes in the Exam format and/or content, the respondents were quite cautious and expressed quite a limited amount of ideas (Figure 9.5), mostly about the Linguistic part of the Exam.

Figure 9.5: Teachers' suggestions of Exam changes

Interview 2	N:	I would not change anything. I would administer this part before [emphasizes] the State Exam. I don't know how it could be done, but it should be done.
	INT:	You mean, State Exam format should be reconsidered.
	N:	Exactly
Interview 3	T:	I would somehow reduce this part. Probably...less questions or questions should be smaller.
	M2:	A multiple choice test? What about a test? Matching could be employed to check terminology

However, when the respondents were asked to choose as many items as they felt necessary from the suggested list of areas⁵⁴, their opinions were much more varied (Figure 9.6).

Figure 9.6: Summary of responses to the question: 'From your current position, what parts, do you think, the exam should consist of?'



⁵⁴ The list of possible assessment areas was designed as a part of the Interview framework on the basis of Literature review (Chapters 3, 4). The teachers (respondents) could add any other part of Exam if they thought it was necessary

The figure above demonstrates that each suggested option was chosen by minimum 3 respondents, although some respondents admitted that if all the parts are included the Examination will be unreasonably long. Nine out of 11 teachers stated that speaking on general topics should be a part of the Final Language Examination, 8 teachers consider Professional (Teacher) writing, Vocabulary (both general and professional) and grammar to be necessary parts of the Exam; 7 people chose Professional listening, Professional reading, Teacher language awareness and Classroom language. The minimal number of respondents (3 out of 11) chose linguistic terminology as a possible area of assessment at the Final Examination.

Ten out of 11 respondents consider the Final Language Examination **obligatory** for all university graduates. Nine people said the Examination should have both written and oral parts, one teacher thinks the examination should be oral and one could not think of any option.

9.3. Needs analysis of teachers of English as a Foreign Language

As described in Chapter 6: Research Methodology, the purpose of the Needs Analysis was to investigate which language skills teachers of English in Russia need for their everyday professional life, and how confident they feel within various areas of the English language. The questionnaire consisted of three questions, with Q1 and Q2 having two subsections each. Q1-2 asked the respondents to rate how often they are involved in the receptive and productive activities listed. Q3 asked how confident teachers feel/felt within different areas of general and teacher English.

Table 9.8 illustrates how often the respondents are involved in different **listening** activities in English – in and out of the classroom.

Table 9.8: Responses to Q1.1: *Please rate how often you are (were) involved in the following listening activities*⁵⁵

	Never	Seldom (1-2 times a year)	Sometimes (3-10 times a year)	Often (1-3 times a month or so)	Very often (once or several times a week)
Listening to speakers at conferences, seminars, webinars	12 (11%)	56 (52.3%)	33 (30.8%)	4 (3.7%)	1 (0.9%)
Listening to pupils/students performing in the classroom	1 (0.9%)	0	1 (0.9%)	9 (8.4%)	96 (89.7%)
Listening to students at teaching practice reflecting on their lessons	35 (32.7%)	21 (19.6%)	12 (11.2%)	15 (14%)	1 (0.9%)
Watching TV, listening to radio, including on-line, for information	13 (12.1%)	29 (27.1%)	21 (19.6%)	14 (13.1%)	29 (27.1%)
Listening to recordings to coursebooks	2 (1.9%)	0	8 (7.5%)	24 (22.4%)	72 (67.3%)
<i>N = 107</i>					

As can be seen from the table above, activities performed most often either take place in the classroom (listening to pupils' performance), or somebody else's classroom (observing English lessons), or during lesson planning (listening to recordings to coursebooks). Not many respondents listen to speakers at conferences and other events (52% do it 1-2 times a year and 11% never do it), listen to students reflecting on their lessons at teaching practice (about 20% of respondents do it once or twice a year whereas nearly 33% of teachers never do it). As far as watching TV/listening to the radio is concerned, there is approximately the same number of those who do it rarely or never do it (27% and 12% correspondingly) and those who do it often or very often (13% and 27% of respondents).

Responses to the question about **reading** activities performed by the teachers were classified in the same way. The results are presented in Table 9.9.

⁵⁵ The maximum number of responses for each activity in Table 9.8-9.12 is shown in **bold**; lists of activities for Q 1.1-1.4 of Needs Analysis were informed by taxonomies of teacher communicative skills reviewed in Chapter 3 and by the taxonomy of general language skills presented in the Common European Framework of Reference (2001)

Table 9.9: Responses to Q1.2: *Please rate how often you are (were) involved in the following reading activities*

	Never	Seldom (1-2 times a year)	Sometimes (3-10 times a year)	Often (1-3 times a month or so)	Very often (once or several times a week)
Reading ELT literature, including ELT magazines	1 (0.9%)	17 (15.9%)	32 (29.9%)	42 (39.3%)	14 (13.1%)
Reading books/articles on Linguistics/Philology	52 (48.6%)	33 (30.8%)	8 (7.5%)	5 (4.7%)	5 (4.7%)
Reading teacher's books and resource packs for lesson planning	0	3 (2.8%)	15 (14%)	34 (31.8%)	51 (47.7%)
Reading reference materials (dictionaries, grammar books, etc.)	0	0	16 (15%)	48 (44.9%)	35 (32.7%)
Reading student writing for giving feedback	1 (0.9%)	1 (0.9%)	8 (7.5%)	41 (38.3%)	53 (49.5%)
Browsing web-sites for teachers	6 (5.6%)	16 (15%)	29 (27.1%)	28 (26.2%)	26 (24.3%)
Reading book catalogues	6 (5.6%)	38 (35.5%)	38 (35.5%)	18 (16.8%)	2 (1.9%)
Reading job adverts	59 (55.1%)	30 (28%)	7 (6.5%)	1 (0.9%)	3 (2.8%)
Reading exam sample papers (including examinations for language teachers)	3 (2.8%)	12 (11.2%)	24 (22.4%)	32 (29.9%)	35 (32.7%)
<i>N = 107</i>					

Similarly to listening activities, reading activities in which the respondents are often involved, are mostly performed either in the classroom (e.g. reading student writing) or for lesson planning (reading teacher's books and resource materials; reading reference materials; studying exam sample papers) and professional development (reading ELT literature). Browsing web-sites for teachers is another quite popular activity: although the biggest number of respondents chose the 'sometimes' category (about 27%), 26% and 24% of respondents stated they do it 'often' and 'very often'.

The least frequent activities are reading books/articles on Linguistics (nearly 49% of respondents never to this); reading job adverts (55% never do it) and reading book catalogues (35.5% of respondents do it rarely and 35.5% do it sometimes only).

The question about **speaking** activities which teachers of English perform in and out of the classroom yielded the responses presented in Table 9.10.

Table 9.10: Responses to Q2.1: *Please rate how often you are (were) involved in the following speaking activities*

	Never	Seldom (1-2 times a year)	Sometimes (3-10 times a year)	Often (1-3 times a month or so)	Very often (once or several times a week)
Making a report/presentation at conferences/ seminars	38 (35.5%)	38 (35.5%)	21 (19.6%)	7 (6.5%)	1 (0.9%)
Making a report on-line	90 (84.1%)	8 (7.5%)	4 (3.7%)	1 (0.9%)	3 (2.8%)
Storytelling	54 (50.5%)	22 (20.6%)	13 (12.1%)	5 (4.7%)	1 (0.9%)
Giving feedback on student performance	1 (0.9%)	2 (1.9%)	6 (5.6%)	12 (11.2%)	82 (76.6%)
Giving feedback on observed lessons	14 (13.1%)	30 (28%)	34 (31.8%)	17 (15.9%)	9 (8.4%)
Explaining language items to pupils/students	1 (0.9%)	1 (0.9%)	1 (0.9%)	12 (11.2%)	91 (85%)
Giving instructions for activities	0	0	3 (2.8%)	8 (7.5%)	92 (86%)
Maintaining discipline in the classroom	1 (0.9%)	2 (1.9%)	4 (3.7%)	23 (21.5%)	73 (68.2%)
Talking to colleagues at conferences, seminars, etc. (including asking questions)	25 (23.4%)	41 (38.3%)	23 (21.5%)	12 (11.2%)	5 (4.7%)
<i>N = 107</i>					

The table above demonstrates that there are 4 activities performed frequently by the respondents: giving feedback on student performance (nearly 77% of teachers do it very often and about 11% - often); giving instructions for activities (86% and 7.5% respectively); explaining language items to pupils/students (85% and 11%) and maintaining discipline in the classroom (with 68% of respondents doing this very often and 21% often). Compared with receptive skills of listening and reading, the range of speaking skills that teachers of English require in their everyday professional life is quite narrow. It might be explained by the wider context – in Russia, where English has a status of a foreign language, exposure to oral communication through English, whether real-life or online (virtual) is quite limited.

Responses to Q2.2 about the **writing** activities teachers find themselves involved in are presented in Table 9.11.

Table 9.11: Responses to Q2.2: *Please rate how often you are (were) involved in the following writing activities*

	Never	Seldom (1-2 times a year)	Sometimes (3-10 times a year)	Often (1-3 times a month or so)	Very often (once or several times a week)
Giving written feedback on student work	3 (2.8%)	6 (5.6%)	14 (13.1%)	36 (33.6%)	45 (42.1%)
Writing references (for colleagues, students)	35 (32.7%)	33 (30.8%)	24 (22.4%)	8 (7.5%)	5 (4.7%)
Writing lesson plans	0	1 (0.9%)	2 (1.9%)	23 (21.5%)	77 (72%)
Designing teaching materials	2 (1.9%)	1 (0.9%)	26 (24.3%)	47 (43.9%)	29 (27.1%)
Making presentations (for lessons, conferences, etc.)	20 (18.7%)	36 (33.6%)	26 (24.3%)	13 (12.1%)	1 (0.9%)
Writing letters to ELT journals	93 (86.9%)	5 (4.7%)	2 (1.9%)	1 (0.9%)	0
Posting comments on ELT sites	64 (59.8%)	28 (26.2%)	8 (7.5%)	2 (1.9%)	4 (3.7%)
Writing articles	64 (59.8%)	24 (22.4%)	11 (10.3%)	4 (3.7%)	3 (2.8%)
Writing formal letters (e.g. to ELT or other journals, teaching institutions, etc.)	61 (57%)	21 (19.6%)	11 (10.3%)	6 (5.6%)	5 (4.7%)
<i>N = 107</i>					

Table 9.11 shows that there are three activities which are done by the majority of respondents: writing lesson plans (21.5% of teachers do it often and 72% - very often); designing teaching materials (nearly 44% and 27% correspondingly); giving written feedback on student work (about 34% and 42% of respondents). According to the data presented, the rest of activities in the list are either never done (e.g. writing letter to ELT journals) or done rather infrequently (e.g. writing articles or formal letters).

Another area investigated in the Needs Analysis is **how confident the respondents feel in different areas of the English language**, both in everyday and professional domain. As Table 9.12 below illustrates, the percentage of those who are not too confident is not higher than 10% in each category. Most respondents feel *confident enough* or *very confident*. For some skills the difference between the number of confident and unconfident people is quite big (e.g. listening, reading, writing skills); for some skills (e.g. command of specialized vocabulary) the difference is smaller.

Table 9.12: Responses to Q3: *Please rate how confident you feel in the following areas of the English language*

Areas of English	Extremely unconfident	Not confident enough	Neutral /do not know	Rather confident	Very confident
listening	1 (0.9%)	1 (0.9%)	13 (12.1%)	66 (61.7%)	25 (23.4%)
reading	0	0	4 (3.7%)	45 (42.1%)	57 (53.3%)
writing	0	6 (5.6%)	14 (13.1%)	59 (55.1%)	27 (25.2%)
speaking – accuracy	0	5 (4.7%)	9 (8.4%)	72 (67.3%)	19 (17.8%)
speaking – fluency	0	9 (8.4%)	25 (23.4%)	58 (54.2%)	14 (13.1%)
general vocabulary	0	2 (1.9%)	5 (4.7%)	47 (43.9%)	50 (46.7%)
specialized language teaching (ELT) terminology	0	8 (7.5%)	37 (34.6%)	47 (43.9%)	14 (13.1%)
classroom language	0	2 (1.9%)	18 (16.8%)	44 (41.1%)	42 (39.3%)
grammar	0	2 (1.9%)	9 (8.4%)	62 (57.9%)	33 (30.8%)
ability to explain language items to pupils	0	3 (2.8%)	15 (14%)	60 (56.1%)	19 (17.8%)
pronunciation	1 (0.9%)	0	28 (26.2%)	51 (47.7%)	25 (23.4%)
<i>N = 107</i>					

To check if the level of confidence depends on respondents' experience data was cross-tabulated and the following tables were produced, using SPSS software. Only several skills are presented below to trace if there is any correlation between the level of confidence and the teaching experience of the respondents: one of the receptive skills (listening) and one of the productive skills (*speaking*) were chosen, and also *grammar* and *teacher-specific vocabulary* (Tables 9.13a-d⁵⁶).

Table 9.13a: Cross-tabulation experience*level of confidence in *listening*

	listening						Total
	extremely unconfident	not confident enough	neutral/ do not know	rather confident	very confident	no response	
5 years or less	0	1	3	12	5	0	21
6-10 years	0	0	0	16	4	0	20
11-20 years	0	0	4	23	7	0	34
more than 20 years	1	0	6	15	8	1	31
Total	1	1	13	66	24	1	106

⁵⁶ Tables 9.13a-d are numbered this way because they all present responses to one question 'How confident do you feel in the following areas of the English language?' Data in tables a-d shows level of respondents' confidence in different skills

Table 9.13b: Cross-tabulation experience*level of confidence in speaking

	speaking-accuracy					Total
	not confident enough	neutral/ do not know	rather confident	very confident	no response	
5 years or less	1	3	13	4	0	21
6-10 years	0	1	15	4	0	20
11-20 years	3	3	23	5	0	34
more than 20 years	1	2	20	6	2	31
Total	5	9	71	19	2	106

	speaking-fluency					Total
	not confident enough	neutral/ do not know	rather confident	very confident	no response	
5 years or less	2	4	9	6	0	21
6-10 years	1	3	12	4	0	20
11-20 years	3	11	17	3	0	34
more than 20 years	3	7	20	0	1	31
Total	9	25	58	13	1	106

Table 9.13c: Cross-tabulation experience*level of confidence in grammar

	grammar					Total
	not confident enough	neutral/ do not know	rather confident	very confident	no response	
5 years or less	1	2	12	6	0	21
6-10 years	1	0	12	7	0	20
11-20 years	0	4	21	9	0	34
more than 20 years	0	3	17	10	1	31
Total	2	9	62	32	1	106

Table 9.13d: Cross-tabulation experience*level of confidence in ELT terminology

	ELT terminology					Total
	not confident enough	neutral/ do not know	rather confident	very confident	no response	
5 years or less	1	8	11	1	0	21
6-10 years	0	6	11	3	0	20
11-20 years	3	11	12	8	0	34
more than 20 years	4	11	13	2	1	31
Total	8	36	47	14	1	106

As can be seen from Tables 9.12 and 9.13a-d, the respondents' teaching experience does not seem to influence the results obtained. For example, the majority of responses for *listening* in Table 9.12 (nearly 62%, or 66 responses out of 106) demonstrate that teachers feel rather confident about it irrespective of their lengths of experience. Table 9.13a, which presents the cross-tabulated data, is in good keeping with Table 9.12: in each category of responses the majority have still chosen the 'rather confident' option. The same tendency is observed in Tables 9.13b-d.

9.4. Needs analysis of the Final year university students

This **Needs analysis**⁵⁷ was performed online among the **university 4th and 5th year students**. The purpose of this internet survey (see Chapter 6: Research methodology: pp.140-142) was to identify the skills that the students employ at school during their teaching practice and how confident they feel using these skills in and out of the classroom. The list of skills the respondents chose from was similar to the list of activities for practicing teachers of English (see above, Needs Analysis of School Teachers) but included mostly classroom activities and lesson preparation, as these are the activities students are expected to do during their teaching practice. Besides, the frequency with which the activities can be performed was different for the student survey – instead of asking practicing teachers about how many times a year they do this or that activity, the students were asked to mark the number of times they did this or that task (Appendix 10). The data was analysed automatically through Survey Monkey (see Chapter 6: Research methodology: pp.142-144).

Table 9.14 presents the students' responses to the question about their employment of listening and reading skills during teaching practice(s) at school.

⁵⁷

https://ru.surveymonkey.com/MySurvey_EditorFull.aspx?sm=YfMIDUuZuR4gw3us%2bEFKuALL9ZrKcwIL%2fChpPtH0AJ0%3d

Table 9.14: Responses to Q4: *As a trainee teacher at school, how often were you involved in [the following] receptive activities in English?*

	Never or once	Seldom (2-4 times)	Sometimes (5-6 times)	Often (nearly every week)	Very often (nearly every day)
Listening to pupils performing in the classroom	0	0	1 (9%)	0	10 (90.91%)
Observing English lessons	0	1 (9%)	5 (45.45%)	4 (36.36%)	1 (9%)
Listening to recordings to coursebooks	1 (9%)	0	1 (9%)	2 (20%)	6 (60%)
Watching news	1 (9%)	1 (9%)	2 (18%)	0	0
Watching (taking part in) ELT seminars/webinars	5 (45.45%)	1 (9%)	3 (27%)	0	1 (9%)
Reading teachers' books	0	2 (18%)	0	3 (27%)	4 (36.36%)
Reading ELT literature, including magazines	2 (18%)	2 (18%)	2 (18%)	3 (27%)	0
Reading newspapers/magazines, including online	1 (9%)	2 (18%)	1 (9%)	0	0
Reading books (fiction)	2 (18%)	1 (9%)	0	0	0
Browsing web-sites for teachers	1 (9%)	1 (9%)	2 (18%)	2 (18%)	2 (18%)
Reading pupils' written work	0	1 (9%)	1 (9%)	5 (45.45%)	4 (36.36%)
Reading books/articles on Linguistics/Philology	4 (36.36%)	2 (18%)	2 (18%)	1 (9%)	1 (9%)
Reading reference books (grammar, dictionaries, etc.)	0	0	3 (27%)	2 (18%)	5 (45.45%)
Reading book catalogues	2 (18%)	1 (9%)	1 (9%)	2 (18%)	0

Note: for some skills the total number of responses is less than 100% because of missing answers

Similarly to the responses of practicing teachers of English (see part 9.2) the most popular listening and reading activities are those required in the classroom or for lesson planning: *listening to pupils performing in the classroom* (nearly 91% of respondents did it very often); *listening to recordings to coursebooks* (20% of students did it often and 60% - very often); *reading teachers' resource materials* (33% did it often and 44% - very often, though there are 22% of those who seldom were involved in these activities); *reading pupils' written work* (45% and 27% correspondingly). Similarly to more experienced teachers, final year students never or seldom listen to speakers at events and read books/articles on Linguistics.

Table 9.15 presents the trainee teachers' responses to the question about most common activities during their teaching practice that involve speaking and writing skills.

Table 9.15: Responses to Q3 'As a trainee teacher at school, how often were you involved in [the following] productive activities in English?'

	Never or once	Seldom (2-4 times)	Sometimes (5-6 times)	Often (nearly every week)	Very often (nearly every day)
Maintaining discipline	0	0	0	2 (18%)	9 (81%)
Reflecting on conducted classes	0	0	4 (36%)	0	6 (54%)
Retelling/discussing articles	1 (9%)	2 (18%)	2 (18%)	0	5 (45%)
Giving instructions	0	0	0	0	9 (81%)
Explaining language items	0	0	0	2 (18%)	9 (81%)
Giving feedback	0	0	3 (27%)	2 (18%)	6 (54%)
Conducting classes	0	0	0	0	10 (90%)
Writing letters	0	2 (18%)	6 (54%)	0	2 (18%)
Posting comments on ELT sites	0	6 (54%)	0	0	3 (27%)
Designing materials	0	0	2 (18%)	0	8 (72%)
Correcting written work	0	0	0	0	9 (81%)
Writing lesson plans	0	0	0	0	11 (100%)
<i>N = 11</i>					

As can be seen, the most required speaking skills employed in the classroom are: *conducting classes* (90% of respondents stated they did it very often); *giving feedback on pupils' performance* (54% of students did it very often and 18% - often); *explaining language items to students* (81% of respondents did it very often); *giving instructions for activities* (81% did it very often); *writing lesson plans* (100% of respondents did it very often).

Similarly to the Needs Analysis for practicing teachers, final year students were asked **how confident they felt** in different areas of English, in and out of the classroom. These responses are presented in Table 9.16.

Table 9.16: Responses to Q5: *Please rate how confident you feel in the following areas of the English language*

	Extremely unconfident	Not confident enough	Neutral / do not know	Rather confident	Very confident
General vocabulary	0	0	0	8 (80%)	2 (20%)
Specialised language teaching (ELT) terminology	1 (10%)	0	4 (40%)	5 (50%)	0
Grammar	0	0	0	9 (90%)	1 (10%)

	Extremely unconfident	Not confident enough	Neutral / do not know	Rather confident	Very confident
Pronunciation	0	0	2 (20%)	7 (70%)	1 (10%)
Classroom language	0	0	4 (40%)	4 (40%)	2 (20%)
Ability to explain language items to pupils	0	1 (10%)	1 (10%)	3 (30%)	1 (10%)
Listening	0	0	0	10 (100%)	0
Reading	0	0	0	7 (70%)	3 (30%)
Speaking - accuracy	0	0	1 (10%)	9 (90%)	0
Speaking - fluency	0	0	0	8 (80%)	2 (20%)
Writing	0	0	0	9 (90%)	1 (10%)
<i>N = 10</i>					

The respondents feel rather confident or very confident in many skills and areas – *grammar, pronunciation, listening, reading, speaking* (both accuracy and fluency); *writing*. In contrast to their more experienced colleagues, practicing teachers, the final year students feel less confident with *Classroom Language* (40% found it difficult to answer the question whereas 60% still feel confident) and *ability to explain language items to students* (nearly 17% stated they were not confident enough; 17% did not know).

When the data from Needs Analysis of school teachers and Final year university students was generated (as shown above), the next step was comparing the skills that those 2 groups of respondents (experienced and beginner teachers) require in and out of the language classroom. The skills for each group of respondents were put in the order of priority, starting with those most frequently employed. Tables 9.17a,b – 9.20a,b demonstrate how often listening, reading, speaking and writing skills are required by experienced teachers of English and future teachers, final year university students.

Table 9.17a: Listening skills teachers of English in Russia require

Listening sub-skills (in the order of priority)	Percentage of respondents who require it very often (once or several times a week)
Listening to pupils in the classroom	89.7%
Listening to recordings to coursebooks	67%
Watching TV, listening to news	27%
Observing English lessons	2.8%
Listening to students at teaching practice	0.9%
Listening to speakers at conferences	0.9%

Table 9.17b: Listening skills student teachers of English in Russia require during teaching practice

Listening sub-skills (in the order of priority)	Percentage of respondents who require it very often (nearly every day)
Listening to pupils in the classroom	90.9%
Listening to recordings to coursebooks	60%
Watching/taking part in ELT seminars	10%
Observing English lessons	9%
Watching news	0%

Table 9.18a: Reading skills teachers of English in Russia require

Reading sub-skills (in the order of priority)	Percentage of respondents who require it very often (once or several times a week)
Reading student writing for giving feedback	49.5%
Reading Teacher's Books and resource packs for lesson planning	47.7%
Reading exam sample papers	32.7%
Reading reference materials (dictionaries, grammar books, etc.)	32.7%
Browsing web-sites for teachers	24%
Reading ELT literature, including magazines	13%
Reading books/articles on Linguistics	4.7%
Reading job adverts	2.8%
Reading book catalogues	1.9%

Table 9.18b: Reading skills student teachers of English in Russia require during teaching practice

Reading sub-skills (in the order of priority)	Percentage of respondents who require it very often (nearly every day)
Reading reference books	50%
Reading Teacher's Books	44.4%
Reading pupils' written work	27.3%
Browsing web-sites for teachers	20%
Reading books and articles on Linguistics/Philology	10%
Reading ELT literature	0%
Reading book catalogues	0%
Reading exam sample papers	0%

Table 9.19a: Speaking skills teachers of English in Russia require

Speaking sub-skills (in the order of priority)	Percentage of respondents who require it very often (once or several times a week)
Giving instructions to activities	86%
Explaining language items to students	86%
Giving feedback on student performance	76.6%
Maintaining discipline in the classroom	68%
Giving feedback on observed lessons	8.4%
Talking to colleagues at conferences	4.7%

Speaking sub-skills (in the order of priority)	Percentage of respondents who require it very often (once or several times a week)
Making a report online	2.8%
Making a report/presentation at seminars/conferences	0.9%
Storytelling	0.9%

Table 9.19b: Speaking skills student teachers of English in Russia require during teaching practice

Speaking sub-skills (in the order of priority)	Percentage of respondents who require it very often (nearly every day)
Conducting classes	100%
Explaining language items to pupils	81%
Giving instructions to activities	81%
Maintaining discipline	81%
Giving feedback on pupil performance	54%
Reflecting on conducted lessons	54%
Retelling what you read/heard	45%

Table 9.20a: Writing skills teachers of English in Russia require

Writing sub-skills (in the order of priority)	Percentage of respondents who require it very often (once or several times a week)
Writing lesson plans	72%
Giving written feedback on student work	42%
Designing teaching materials	27%
Writing references (to colleagues, students)	4.7%
Writing formal letters	4.7%
Posting comments on ELT sites	3.7%
Writing articles	2.8%
Writing letters to ELT journals	0%

Table 9.20b: Writing skills student teachers of English in Russia require during teaching practice

Writing sub-skills (in the order of priority)	Percentage of respondents who require it very often (nearly every day)
Writing lesson plans	100%
Correcting written work, giving written feedback	81%
Designing teaching materials	72%
Posting comments on ELT web-sites	27%
Writing formal letters	18%

As the above demonstrates, for both groups of respondents the skills that are employed in the language classroom or for lesson preparation are those most frequently required.

Both experienced and student teachers quite rarely use English outside school that can be explained by the context of teaching and learning foreign languages in Russia.

Whilst **Survey 1** provided information on the content and format of the Final Language Examination, teacher interviews cast some light on how practicing teachers of English see the Examination and whether they find it relevant to their job. **Surveys 3 and 4** provided the data about the skills that teachers most widely employ, and activities they deal with. All collected data will be used further in this work to compare what is currently assessed at the Exam with what was singled out and highlighted by the teachers. This will result in further suggestions of possible changes in the Exam content, format and administration.

Chapter 10

Discussion of findings on Research Question 2: How relevant is the Examination content to the language needs of practicing English teachers? What are these needs?

This chapter discusses the three sets of data presented in Chapter 9. The first set, obtained from Survey 1, is a set of ‘factual’ data from Exam designers and examiners on what the Final Language Examination consists of and what knowledge areas and skills it aims to test.

The 2nd set of data, the responses to Surveys and 3, 4 for English teachers, presents the language activities in which teachers are involved in the process of teaching English. This data is essential for discussing validity and authenticity of the current Examination.

The 3rd set of data also deals with the Examination content and format but is based on opinions rather than facts, when teachers of English are asked to reflect on how relevant is what the Examination assesses to what they require every day in the language classroom.

Chapter 10 consists of two parts. First, the chapter reflects on the obtained data and compares facts and opinions provided by different stakeholders. Then some concerns relating to the Exam content and format are discussed.

10.1. Final Language Examination content and format as seen by the stakeholders

To speculate on the content and format of the current Examination, the data from 2 surveys was summarised. In Survey 1 for examiners and Exam designers the respondents were asked what was in the focus of assessment at the Final Language Examination (Chapter 9: Tables 9.1-9.3). Quite predictably, the responses outlined 3 foci: linguistic knowledge, listening, and reading skill, as defined by the Final Examination Syllabus (2010: 36-40). Judging by the responses, the content of

assessment is predominately general English, with teacher English being almost out of the Exam focus. As far as the assessment format is concerned, no detailed information was traced in the Exam Syllabus: data from Survey 1 for examiners provided more evidence to the fact that neither task types employed in the Exam, nor their number and distribution within Exam sections were described in detail in documents or agreed upon within the team of Exam designers.

Analysis of the obtained data revealed some issues about the content and format of the current Final Language Examination. First, a substantial linguistic task assesses knowledge about language; no tasks or elements have been traced that would assess graduates' ability to *teach* language. Second, the reading section assesses essential reading for detailed understanding skills in general English, whilst teacher-related content is represented scarcely in texts for reading. Similarly to reading, the listening section assesses listening for detailed understanding in general English only. Speaking is represented through prepared monologue. Teacher-related communicative skills, including Classroom English, seem to be out of the assessment focus. In addition to these, no rationale for choosing concrete text types or guidelines on content representation have been found either in the Exam Syllabus or in the collected empirical data.

To discuss sufficiency and appropriacy of the content and format of the current Examination, opinion of 'neutral' stakeholders was sought. In teacher interviews (Chapter 9; Appendix 11, 12) the respondents were asked to reflect upon usefulness and appropriacy of the current Exam tasks from their position of practicing teachers of English. None of the participants expressed much doubt about appropriacy of reading and listening tasks, although 2 people admitted that they never had to do such tasks in real life – they do read a lot in English but they never had to retell what they read *in English* (Chapter 9: Figures 9.2–9.4). This situation can be explained by the context of teaching and learning foreign languages in Russia⁵⁸: very often teachers of FL can only use it in the classroom or at special events, but not in their everyday life or communicating with colleagues. In-service courses that every teacher must take every

⁵⁸ The statement at this stage of research can be supported by anecdotal evidence only. No statistical data was collected to support this statement; it can be viewed as a direction for further research

5 years, quite rarely aim to support and upgrade teachers' language skills: lectures and seminars are usually delivered in Russian, and even the TESOL methods course or a course in innovations in language teaching are conducted in Russian. Sometimes language teachers attend lectures on general and even political issues that, as one of the teachers said 'can be really interesting to listen, but have nothing to do with my job'. An attempt was made in 2000-2007 to organise In-service courses for college teachers of English within the British Council PRESET project (see Chapter 2). The courses that were held for 5 years and involved more than 100 college teachers of English, included several modules conducted in English and a Language Development module done by a native speaker (invited by the BC). In addition to these, participants did small-scale research within an ELT area of their interest. In 2007, when majority of colleges stopped training English language teachers, the in-service courses were closed, too.

Although the respondents' overall opinion of reading and listening tasks was positive, some interviewees suggested widening the range of texts for reading and listening and making them more professionally-related (Chapter 9: Table 9.7) and expressed their concern about the speaking component in the Exam tasks being limited to practically one sub-skill (Chapter 9: Figure 9.4, Table 9.6). One of the interviewees was concerned about fully prepared retelling of a text being the only speaking task in Exam Tasks 2 and 3. She referred to the final language examination she had taken when graduating from the teacher training college (Chapter 2; Appendix 3) where graduates were supposed to react spontaneously on verbal and visual input. Another concern expressed in interviews was lack of professional dimension in speaking tasks. This can be treated as a consequence of choice of reading and listening texts, when Exam takers have to react to non-professional content.

The part of the Exam that elicited quite contradictory opinions was Task 1: Linguistic questions. Many interviewees considered this task appropriate and useful for teachers (Chapter 9: Figure 9.1). At the same time, many teachers with different levels of experience were a bit apprehensive of whether such a task was useful and whether it assessed what teachers of English really need in the classroom (Chapter 9: Figure 9.1, Table 9.5). None of the respondents was in doubt whether linguistic knowledge was essential for language teachers. Nevertheless, they doubted the necessity to reproduce that amount of information with a perspective of never requiring it in teaching. Some

teachers, without mentioning the concept, referred to Teacher Language Awareness (Chapter 3) described by Bolitho (2003) as an ability of [trainee] teachers:

'to analyse language, to apply different strategies for thinking about language (analogizing, contrasting, substituting) in order to be able to plan lessons, to predict learners' difficulties, to answer their questions, and to write and evaluate materials' (2003: 255)

The major point made by some teachers is lack of connection between a substantial amount of linguistic knowledge that university graduates are supposed to demonstrate and its limited applicability at school due to this knowledge 1) not being a part of school FL curriculum and 2) not leading directly to an ability to teach language items to different age and ability groups. When asked about possible ways of changing the current situation, the respondents suggested a range of steps – from making questions 'smaller in scope' and changing the task types (there was a suggestion to transform the questions into a written test) to removing the task from the Final Exam (Chapter 9: Figure 9.5).

Having different amount of experience, all respondents felt some discrepancy between what teachers do in the classroom and what is assessed at the Final Language Exam. Major points of criticism in the current Exam, according to the interviewees, were:

- too extended linguistic task (Task 1) that does not assess what teacher employ in the classroom;
- a limited range of tasks for listening and reading;
- prepared monologue in both Task 2 and Task 3 with a very limited element of spontaneity.

Apart from asking school teachers about the relevance of the current Final Language Examination to their job, the focus of the Exam was compared to the activities that teachers deal with in and out of the language classroom (part 9.3-4).

As stated previously, the data collected in this study aims to provide a clearer and more precise picture of what teachers of English as a Foreign Language in Russia need in and out of the language classroom in terms of knowledge and skills. Needs analysis of language teachers at schools, colleges and universities yielded a battery of activities and skills that teachers require seldom, often and very often (Chapter 9: Tables 9.8-9.12). Then the activities were put in the order of priority, with the most required ones

being on top of the list (Tables 9.17a-9.20a). The collected data demonstrates that the respondents are most often involved in conducting classes (including giving instructions, explaining language items, giving feedback, etc.), lesson planning and marking student work, reading teacher resources and ELT literature. The respondents quite rarely do presentations or reports, talk to colleagues in English, write articles or post comments on ELT sites, or read outwith the ELT area. Such results were, to a certain extent, predictable for the Russian context of teaching and learning foreign languages. Russian is the only official language of the Russian Federation⁵⁹, with English, French, German, Spanish, Chinese and other languages having the status of foreign languages. Therefore, Russian is the official language for documentation, including all school documentation, official events, mass media, web-sites based in Russia, etc. Often enough, teachers of English and other FL do not have an opportunity to use English as a language of communication, whether communication is face-to-face or virtual. The Internet has brought considerable changes in terms of availability of sources, so teachers gained more access to texts for reading and listening – for themselves and their students. As far as speaking and writing are concerned, the Internet has also offered some extra opportunities, but still has not changed the balance of skills that teachers employ.

The results of the Needs Analysis of practising teachers were compared with the results of Needs Analysis for trainee teachers of English (Chapter 9, part 9.4: Tables 9.14, 9.15). No considerable difference was found between what experienced teachers and trainee teachers require – for both categories the skills employed in the language classroom were prevailing. Some difference was observed in reading – trainee teachers read more literature on Linguistics and Philology and less ELT sources. This might be explained by the fact that final year students do a lot of linguistic courses and prepare for their Final Exam at university, where linguistic knowledge is crucial, whilst practising teachers are more interested in practical issues of teaching languages. In addition to this, trainee teachers get guidance from their mentors, including advice on

⁵⁹ There are territories with other national (minority) languages functioning as second (after Russian) official languages (Tatarstan, Yakut Republic and some others), but this does not change the status of English, German, French and other languages that are foreign

lesson planning, resources to be used, etc., so they might not need to read ELT books or magazines in search of the information they require.

As this research focuses on the assessment of *language* development of trainee teachers, the issue that raises quite a number of questions is identifying language skills that are involved in those activities. It is fully understood that identifying language skills in professional activities is a complicated task due to their highly integrated nature. Thus, everything that teachers do in the classroom requires professional knowledge, be it linguistic knowledge, TLA (Chapter 3) or knowledge of psychology and/or teaching methods. For example, reading a Teacher’s Book for lesson planning would involve, apart from reading skills, a command of ELT terminology; explaining a language item would require linguistic knowledge and also an ability to foresee possible learner difficulties, and an ability to find suitable ways of explanation.

In this research, the major purpose of English teacher Needs Analysis is seen as identifying most common activities for teachers and mapping out a range of skills each activity involves (Table 10.1). This would allow, on the one hand, for a detailed comparison of the current Exam foci with teacher needs and, on the other hand, for providing balanced and adequate suggestions for possible changes in the Exam content and format. The content and layout of the table below were informed by the ICELT exam specifications, where for each language activity specifications present ‘lead’ language skills and supporting skills required for successful task fulfilment. A similar job was done within research projects on teacher language reviewed in Chapter 3 (pp.39-42).

Table 10.1: Involvement of professional knowledge and communicative skills in teacher activities in and out of the language classroom

Activity	What it involves						
	Linguistic knowledge	TLA	Listening	Reading	Speaking	Writing	Other skills/ knowledge
Listening to pupils in the classroom	✓	✓	✓				
Listening to recordings to coursebooks		✓	✓				lesson planning
Watching TV, listening to news			✓				intercultural awareness

Activity	What it involves						
	Linguistic knowledge	TLA	Listening	Reading	Speaking	Writing	Other skills/ knowledge
Listening to trainee teachers at teaching practice		✓	✓				pedagogical knowledge, TESOL
Listening to speakers at conferences		✓	✓	✓			
Reading student writing	✓	✓		✓			
Reading Teacher's Books	✓	✓		✓			TESOL
Reading reference materials	✓	✓		✓			
Browsing ELT web-sites	✓	✓		✓			TESOL
Reading ELT literature		✓		✓			TESOL
Reading book catalogues				✓			
Giving instructions for activities		✓			✓		Classroom English
Explaining language items	✓	✓			✓		Classroom English
Giving feedback on student performance		✓	✓	✓	✓	✓	
Maintaining discipline					✓		Classroom English
Talking to colleagues			✓		✓		TESOL
Making a report /presentation		✓			✓	✓	TESOL
Storytelling				✓	✓		Classroom English
Writing lesson plans	✓	✓		✓		✓	TESOL
Materials design	✓	✓	✓	✓		✓	TESOL
Writing formal letters						✓	
Writing references						✓	
Posting comments on ELT web-sites		✓		✓		✓	TESOL

The taxonomy of skills and activities above was then compared with the assessment foci in the current Final Language Examination. The results of this comparison support the previously expressed concern: out of a variety of skills and knowledge that comprise teacher language competence and that is required by teacher in and out of the classroom, only a fraction is assessed. This can be explained by several reasons, including those analysed by Grant (1997), McNamara (1997) and Elder (2001): assessment of teacher communicative skills is seen as highly problematic outside a language classroom. Therefore, there will always be a threat to authenticity of assessment tasks for teachers, if an exam is administered at the examination centre,

but not while a teacher is conducting a language lesson. Nevertheless, analysis of the task types employed by national and international examination bodies and skills that these tasks assess (Chapter 4) demonstrates that there is some range to choose from and some options to employ in the Final Language Examination under study with minimal threats to its validity and authenticity.

10.2. Reflection on the obtained data: how relevant is the Exam to teacher language needs?

In the process of evaluation of the current Final Language Examination content, some discrepancy occurs from the very first steps. On the one hand, there is a correlation between what is assessed at the Exam and what students are supposed to demonstrate during the course of studies⁶⁰. The balance of subjects in the curriculum is much similar to the balance of content in the Final Language Examination: a substantial component in theoretical Linguistics, and Practical Course of English, with similar tasks and similar expected performance in each of them. On the other hand, comparison of skills and knowledge areas under assessment at the Final Language Examination with a description of language teacher language competence (Chapter 3; this chapter) demonstrates some mismatch between what should be and what is, in fact, assessed. The assessment focus of the Final Language Exam proved to be much narrower than the range of knowledge and skills a FL teacher requires. To support this argument, another comparison was made: the assessment focus of the Exam was juxtaposed with the list of skills teachers of English in Russia require in and out of the classroom (current chapter, part 10.1). The results are similar to those in the previous comparison: only a fraction of what is required at school is assessed at the Final Language Examination at university. Moreover, the skills under assessment at the Exam were rated by some stakeholders, school teachers of English, as those least frequently used by teachers at school (Chapter 9: Figure 9.2) that might question

⁶⁰ Based on comparison of the Final Exam syllabus and subject syllabi on Practical Course of English and theoretical linguistic disciplines: Theoretical Grammar, Theoretical Phonetics, Lexicology, History of English

authenticity of the current Exam tasks. So, some incongruity of the Exam content is in its correlation with the subjects taught throughout all course of studies and, therefore, quite high content validity, and rather poor connection with the taxonomy of language activities that teachers face in the classroom, i.e. threatened construct validity. As claimed by various authors (e.g. Norris, 2009; Chapelle, 2010; Kane, 2010) construct validity can be quite a problematic concept to define and, therefore, argue. Absence of a commonly accepted structure of language teacher language competence (Chapter 3) could undermine any argument against low construct validity of language tests for language teachers. Construct evaluation of the current Final Language Examination is based on a tentative description of teacher language competence and, therefore, may be treated as not completely reliable. To minimise this, this research suggests a taxonomy of activities and teacher communicative skills based on teacher Needs Analysis (Chapter 9). It is hoped that such a taxonomy can serve a springboard for evaluation of construct validity of the current Examination.

As discussed earlier, there might be various reasons for including quite a limited range of skills in the Exam, with the major one being difficulty in providing valid and authentic assessment tasks. Nevertheless, analysis of publications and national and international experience in language tests for language teachers demonstrated that there are quite a number of assessment techniques that can be employed even in an artificial milieu of exam room, as opposed to real life tasks FL teachers do in the language classroom.

The collected data also disclosed some *contradictions* between the current Final Assessment practices in the studied context and the way the stakeholders see those practices. Attitudes of school teachers of English seem to be far from critical, despite their statements on low relevance of the Exam content to their job. Although many respondents to teacher interviews admitted that they never or very rarely in their professional life have to deal with tasks similar to the Exam tasks, they were sure that Reading and Listening tasks ‘were good/useful’ (Interview 1, 2, 3). The typical responses to the question about the Linguistic part of the Exam were ‘it’s too detailed, we never need it at school, but we must know it’ (Interview 1, 2, 3). School teachers seem to be dominated by the principle ‘it has always been like that’ or ‘if it is done this way it means it has to be this way’. For all participants of teacher interviews (aged

between 25 and 43), the Final Language Examination they took at university was the same or approximately the same. Although this fact was known before this research was started, the influence of teachers' own experience and, possibly, limited awareness of other possible options for assessment were underestimated.

One more contradiction was yielded by the Teacher Needs Analysis (Chapter 9, part 9.3). When asked how confident they felt in various areas of general and professional English, the majority of respondents chose options 'rather confident' and 'very confident' (Chapter 9: Table 9.16). There would be nothing contradictory in those responses if analysis of documents and collection of empirical data from various stakeholders had not disclosed so many weak points in the Final Language Examination and, indirectly, in the programme of studies. Study of the Final Exam Syllabus and some subject syllabi revealed that at least 50% of skills presented in Q3 of Survey 4 '*Please rate how confident you feel in the following areas of the English language*' are neither developed in the course of studies nor assessed at the Final Language Examination. Nevertheless, the majority of the respondents stated they felt 'rather confident' in all areas, with the strongest points being listening and reading. There might be several explanations for this fact. The first one is that, despite drawbacks in the model of English teacher development and final assessment, university graduates demonstrate language competence at the level that is sufficient enough to fulfil professional tasks, and if they experience some gaps in their skills, vocabulary or other areas, those gaps can be bridged quite easily through self-studies or other means. Another explanation that might be offered is that the responses to the question were based solely on the respondents' perceptions of the situation. Respondents' self-evaluation is a valid factor, but no other forms of evaluation of language proficiency were applied in this study, and were not supposed to be applied. If the respondents were offered some tasks to do in and/or out of the classroom in addition to the self-evaluation question, the results might have been different, or might have stayed the same.

One of the *concerns*, not directly related to the Final Language Assessment, seems to be incongruity of the situation under study that manifests at 2 levels. First, it is discrepancy between what is assessed at the Final Language Examination for future language teachers and what language teachers need in and out of the classroom in

terms of knowledge and language skills. Some disagreement was also observed between the current Final Assessment foci and the structure of language teacher language competence as understood by researchers and modern examination bodies (e.g. Cambridge ESOL, ETS). Second, it is interdependence of content and format of Final Language Examination and FL teacher training curriculum that makes impossible changes at one level only. Thus, changes in the Final Language Examination might be considered as quite unlikely without changes of the programme of study.

Findings on Research Question 3: What are the strengths and weaknesses of the current Final Language Examination? What changes, if any, might be required?

This chapter presents the findings on the Examination strengths and weaknesses as seen by the Exam developers, administrators and examiners with different amounts of experience. The data was obtained through Survey 1 which involved the respondents in filling out a specially designed questionnaire. Questions 30-35 aimed at getting the respondents' opinion on the strong and weak points of the Exam and possible areas of change.

Q30 of Survey 1 asked the respondents if they feel satisfied with the current form of the Final Language Examination. The answers were cross-tabulated according to the respondents' level of experience in the Examination.

Table 11.1: Responses to Q30: *What do you feel about the exam?*

Respondents' experience in the Final Language Examination	Answers given by the respondents					Total
	completely satisfied	satisfied with the content only	neutral	satisfied with the format only	completely dissatisfied	
less than 2 years	0	0	1	0	0	1
2-5 years	0	0	3	2	0	5
6-10 years	0	0	3	0	1	4
more than 10 years	1	1	4	0	4	10
Total	1	1	11	2	5	20

As can be seen in the table above, most people feel neutral about the Examination, i.e. find it difficult to identify what exactly they are happy/unhappy about. There is one person with more than 10 years of experience in the Exam who is satisfied both with Exam content and format and 5 people with more than 6 years of experience who, on the contrary, are completely dissatisfied with the existing situation. Two respondents with 2-5 years of experience consider the Exam format adequate. Attitudes seem to change with experience – the most dissatisfied group of respondents is the most experienced one. At the same time, it is the most experienced respondents who feel either completely satisfied (1 person) or satisfied with the content (1 person), so the

most experienced group demonstrates a spread of opinions, whereas other groups seem to be more homogeneous.

Apart from overall impression on the Final Language Examination, the respondents were asked what they saw as **advantages** of the Exam. In Q31, they were offered statements to agree/disagree with (Tables 11.2a-f). Tables 11.2a-f were numbered in this way because they all illustrate the respondents' answers to one question: 'Please state what you think about the possible advantages of the examination listed below'. This question was presented in the questionnaire in the form of a table (Appendix 7, Q31) that contained a series of statements and an 'agree-disagree' scale for each statement. The respondents had to agree/disagree/remain neutral about each statement (6 in number). Tables 11.2a-f present responses to those statements. Some responses were cross-tabulated according to the respondents' experience in the Final Examination to see if their vision of the Exam depends on the amount of time they have been involved in it.

Table 11.2a: Responses to Q31: *Please state what you think about the possible advantages of the examination listed below:*

Respondents' experience in the Final Language Examination	The content of questions fits the subject syllabi and the State Standards			Total
	strongly agree	agree	neutral	
less than 2 years	0	1	0	1
2-5 years	1	3	1	5
6-10 years	0	3	1	4
more than 10 years	0	2	8	10
Total	1	9	10	20

As can be seen from the table above, 9 out of 20 respondents agreed to the statement whereas 10 were neutral. A tendency can be observed for less experienced examiners to agree with the statement, with their more experienced colleagues remaining neutral. The difference in opinions might be explained by a very vague content of the Standards (Chapter 5), so when asked, the respondents either chose 'agree', meaning a correlation between the Exam and very general state requirements or 'neutral', meaning that the Standards are a bit too vague to be compared to the concrete Exam tasks and students' expected performance.

Another statement the respondents were offered deals with the Final Language Examination assessing everything graduates need for their job as teachers of English.

The responses showed no big difference in opinions between the groups – in each group there were people who agreed, remained neutral, disagreed and even strongly disagreed. More than 50% of respondents (13 out of 20) do not see the Exam focusing on all necessary skills that future teachers need (Table 11.2b). Still, 3 people consider the Exam content sufficient for future teachers and 4 respondents preferred the ‘neutral’ option.

Table 11.2b: Responses to Q31: *Please state what you think about the possible advantages of the examination listed below*

		The exam assesses everything the graduates will need in the future				Total
		agree	neutral	disagree	strongly disagree	
experience	less than 2 years	0	0	0	1	1
	2-5 years	1	1	2	1	5
	6-10 years	1	0	3	0	4
	more than 10 years	1	3	5	1	10
Total		3	4	10	3	20

The majority of respondents (16 out of 20) consider the Final Language Examination a good opportunity to listen to students (Table 11.2c) and, as it might be presupposed, to get a detailed impression on student performance. It is important, as the Final Language Examination is often the last official language examination English teachers take in their professional life. In-service courses for teachers do not include language assessment and, if some other forms assessment are presupposed, it is usually a group or an individual project, done in Russian. In their professional careers, teachers do need to conduct demo classes every 5 years, that is a good opportunity to demonstrate professional skills, including language skills, but teachers are not supposed to take official language examinations unless they want to (see Chapter 2).

Table 11.2c: Responses to Q31: *Please state what you think about the possible advantages of the examination listed below: Examiners have a good opportunity to listen to candidates and ask questions*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
strongly agree	1	5%
agree	15	75%
neutral	4	20%
Total	20	100%

At the same time, as Table 11.2d demonstrates, 10 respondents disagree and 4 people strongly disagree that the Exam takers are in **equal conditions**.

Table 11.2d: Responses to Q31: *Please state what you think about the possible advantages of the examination listed below: Equal conditions are created for all candidates*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
strongly agree	1	5%
neutral	5	25%
disagree	10	50%
strongly disagree	4	20%
Total	20	100%

Familiarity of Final Language Exam format is viewed as an advantage by 17 out of 20 respondents.

Table 11.2e: Responses to Q31: *Please state what you think about the possible advantages of the examination listed below: Everybody is used to the format, so there is no difficulty in organizing it [the Final Exam]*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
strongly agree	7	35%
agree	10	50%
neutral	2	10%
Total	19	95%
Answer missing	1	5%
Total	20	100%

Familiar formats might lead to easier **administration**: 8 respondents agree with it but 11 remain neutral: there are no specially appointed Exam administrators, so all responsibilities – from listening to students to handing out Exam tasks are performed by examiners (Chapter 5).

Table 11.2f: Responses to Q31: *Please state what you think about the possible advantages of the examination listed below: The examination is easy to administer*

	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
strongly agree	1	5%
agree	7	35%
neutral	11	55%
strongly disagree	1	5%
Total	20	100%

After reflecting on the advantages of the current form of the Final Language Examination, the participants were offered a list of issues that might arise at examinations. This list was compiled as a result of literature review on language testing. Q32 asked if the issues presented in the questionnaire were typical of the Final Language Exam under consideration; Q33 asked if the respondents considered those issues problematic. Tables 11.3a-g⁶¹ present the data obtained from both questions.

Table 11.3a: Responses to Q32: *Please state if the following can happen in your situation (your department); and Q33: Please state if you think the following is a problem which needs to be solved*

Not all core skills are assessed		
	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Does this happen?		
always happens	15	75%
sometimes happens	3	15%
never happens	2	10%
Total	20	

⁶¹ Tables 11.3a-g were numbered in the same way as tables 11.2a-f. They all present the responses to one question (Appendix 7, Q32) where the respondents agreed/disagreed to a series of statement. Each table (a-g) illustrate responses to one statement

Not all core skills are assessed		
Do you consider it as a problem?		
strongly agree	7	35%
agree	10	50%
neutral	2	10%
disagree	1	5%
Total	20	

As can be seen, the majority of respondents (15 out of 20) agree with the statement saying that the situation always takes place, with only 2 people choosing the ‘never happens’ option. As a consequence, 17 people consider this non-sufficient amount of assessed skills as a problem to be solved. The data presented in Table 11.3a, in some way, may be viewed as logical development of the results discussed above (Table 11.2b) – the Final Language Examination tends not to assess the skills that graduates will need in their future job as teachers of English. The responses to the next two sub-questions (Tables 11.3b-c) seem to support this concern.

Table 11.3b: Responses to Q32: *Please state if the following can happen in your situation (your department);* and Q33: *Please state if you think the following is a problem which needs to be solved*

Mostly knowledge is assessed, not skills		
	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Does this happen?		
always happens	5	25%
sometimes happens	13	65%
don't know	1	5%
No answer	1	5%
Total	20	
Do you consider it as a problem?		
strongly agree	1	5%
agree	14	70%
neutral	4	20%
disagree	1	5%
Total	20	

Focus on assessment of knowledge might lead, in some cases, to learning by heart the content of coursebooks and reproducing it at the Exam (Table 11.3c):

Table 11.3c: Responses to Q32: *Please state if the following can happen in your situation (your department);* and Q33: *Please state if you think the following is a problem which needs to be solved*

Answers learnt by heart		
	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Does this happen?		
sometimes happens	17	85%
never happens	2	10%
don't know	1	5%
Total	20	
Do you consider it as a problem?		
agree	13	65%
neutral	5	25%
disagree	2	10%
Total	20	

The next set of sub-questions in Q32-33 dealt with the Final Exam reliability. All respondents indicated that Exam-takers are not always in equal conditions, and the situation is seen as problematic (Table 11.3d).

Table 11.3d: Responses to Q32: *Please state if the following can happen in your situation (your department);* and Q33: *Please state if you think the following is a problem which needs to be solved*

The students are not in equal situations		
	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Does this happen?		
always happens	11	55%
sometimes happens	9	45%
Total	20	
Do you consider this a problem?		
strongly agree	5	25%
agree	14	70%
neutral	1	5%
Total	20	100%

As can be seen, when answering the question about unequal conditions for students being a problem, 19 respondents consider this a problem to be solved, whereas one remains neutral.

A very sensitive area that caused respondents' concern is **cheating**. Cheating is a phenomenon that examiners and exam administrators sometimes find difficult to acknowledge as they feel that by doing so they acknowledge their own helplessness or even involvement in the process. It should be noted that in this situation, the format of the Final Language Examination, in some way, provokes cheating – it is highly possible for Question 1 (Linguistic knowledge) where students are expected to speak about a problem from the Exam card they get, and very often it turns into reproducing chapters from textbooks. For Reading and Listening tasks cheating is unlikely – there is neither a chance nor a source for cheating. In any case, cheating was considered a problem by 16 out of 20 respondents with one person disagreeing.

Table 11.3e: Responses to Q32: *Please state if the following can happen in your situation (your department);* and Q33: *Please state if you think the following is a problem which needs to be solved*

Cheating		
	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Does this happen?		
sometimes happens	18	90%
never happens	2	10%
Total	20	
Do you consider this a problem?		
strongly agree	3	15%
agree	13	65%
neutral	3	15%
disagree	1	5%
Total	20	

Another issue that the respondents consider as 'always' or 'sometimes' happening is a threat to rater reliability – both inter-rater and intra-rater (Table 11.3f). More than 50% of examiners state the problem occurs from time to time with more people (14 out of 20) thinking inter-rater reliability requires more attention than intra-rater reliability (6 people agreeing and 10 remaining neutral).

Table 11.3f: Responses to Q32: *Please state if the following can happen in your situation (your department)*; and Q33: *Please state if you think the following is a problem which needs to be solved*

Lack of co-ordination between examiners (inter-rater reliability)		
	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
Does this happen?		
always happens	3	15%
sometimes happens	13	65%
never happens	2	10%
don't know	2	10%
Total	20	
Do you consider this a problem?		
strongly agree	4	20%
agree	10	50%
neutral	5	25%
disagree	1	5%
Total	20	
Lack of consistency in each examiner's work (intra-rater reliability)		
Does this happen?		
sometimes happens	13	65%
don't know	7	35%
Total	20	
Do you consider this a problem?		
agree	6	30%
neutral	10	50%
disagree	4	20%
Total	20	

The situation may be viewed as highly predictable, with absence of descriptors for criteria, no indication on weighting and no guidelines on marking procedures. Moreover, examiners have to apply the same rather schematic list of criteria to different tasks – both to a theoretical linguistic question and a listening/reading task, and to different situations: questions of different difficulty levels, student answers with different amount of examiner intervention. This is seen to be one of the issues of the current assessment system, discussed further in Chapter 12 together with some ways of changing the situation.

Examination materials design was in the focus of another sub-question in Q32-33. The responses obtained demonstrate that the respondents do not see materials design as an expensive process. The linguistic questions can be reviewed every five years

together with revision of the Final Exam Syllabus. Texts for reading are usually taken from free internet resources and task design only presupposes choosing suitable texts (Chapter 5). No materials piloting takes place, no external experts are involved in task moderation (Chapter 7, Tables 7.8-7.10). All these reduce costs of materials design. Table 11.3g presents only the responses to Q32, where examiners agreed/disagreed with the statement ‘*Exam materials design is very expensive and time-consuming*’. None of the respondents saw this as a problem to be solved.

Table 11.3g: Responses to Q32: *Please state if the following can happen in your situation (your department)*

Exam materials design is expensive	Number of questionnaire respondents who gave this answer	Percentage of total number of respondents
always happens	1	5%
sometimes happens	1	5%
never happens	11	55%
don't know	7	35%
Total	20	

As far as the respondents’ vision of **possible changes** in the Examination, the majority (16 out of 20) stated that changes are necessary. The results were then cross-tabulated according to the respondents’ experience.

Table 11.4: Responses to Q34: *Would you like any changes to be introduced?*

		Would you like any changes to be introduced?			Total
		yes	no	I don't know	
experience	less than 2 years	1	0	0	1
	2-5 years	4	0	1	5
	6-10 years	4	0	0	4
	more than 10 years	7	1	2	10
Total		16	1	3	20

As can be seen from Table 11.4, the majority of respondents in each category see changes as necessary. The most experienced group (more than 10 years of experience) is slightly different from others: 1 person said no changes should take place whereas 2 respondents chose the ‘don’t know’ option. This might be explained by their experience – the longer people are involved in the Examination the more advantages and disadvantages they see. Another factor might be convenience and habit: if the Final

Examination has been administered in this content and format for quite a long period of time, the respondents might see no necessity to change something that works.

As far as **types of changes** the respondents would like to see, the opinions were divided as follows:

Table 11.5: Responses to Q35: *What kind of changes would you like to have?*

<i>Possible changes</i>	<i>Respondents' experience</i>			
	less than 2 years	2-5 years	6-10 years	more than 10 years
New parts added	0	0	2	5
Some parts excluded	0	1	1	1
Content changed	1	3	4	6
Different task types used	0	3	4	7
Administration changed	0	0	3	3

As can be seen, respondents with different levels of experiences suggest changing the Examination content (14 out of 20 people), and using different types of tasks (14 out of 20). Less common suggestions are adding some new parts (7 people), excluding some parts (3 respondents), and changing administration (6 people). A tendency can be traced for more experienced respondents (more than 6 years of experiences as examiners or Exam developers) to come out with more suggestions than their less experienced colleagues.

Chapter 12

Discussion of the Foreign Language Examination strengths and weaknesses as seen by different groups of stakeholders

This chapter presents a discussion of the statistical data obtained from the Exam designers and examiners as to how they see the strong and weak points of the Final Language Examination – from its design to its administration. The chapter also discloses some contradictions between the observed situation and the way this situation is perceived by the stakeholders. After reflecting on and summarizing advantages and disadvantages of the current Examination, the Chapter suggests several ways of changing the situation. Potential changes are viewed as essential, desirable and/or optional, and their feasibility is discussed in relation to the context of FL teacher development in Russia and the opportunities this context provides for teacher training programmes.

Chapter 12 concludes by giving an outline of further steps to be taken for the Exam improvement which might lead to some re-shaping of the course of studies and, more specifically, the system of continuous assessment for pre-service trainee teachers of English at university.

12.1. Opinions of Exam developers as a context for possible changes

Whilst school teachers' opinions of the Exam were discussed previously (Chapter 10), the opinions of Exam developers have not yet been considered. This part deals with Exam developers' visions of the Exam's strong and weak points and their suggestions for changes. This analysis aims to provide a context for further discussion of changes.

When asked about their overall impressions of the Exam, more than a half of the respondents felt neutral, with 1 person completely satisfied and 5 people completely dissatisfied (Chapter 11: Table 11.1). The major advantages of the Exam, as seen by its developers and examiners, are:

- the Exam fits the subject syllabi and requirements of the State Standards (Table 11.2a);
- it provides a good opportunity for examiners to listen to Exam takers (Table 11.2c);
- everybody (both examiners and Exam takers) is used to the format of the Exam (Table 11.2c).

At the same time, the respondents felt quite doubtful about the Exam with respect to:

- assessing all that graduates need (Table 11.2b);
- creating equal conditions for all exam takers (Table 11.2d);
- ease of administration (Table 11.2f).

When speculating on potential Exam threats⁶² and applying them to their situation and the Examination under study, the respondents admitted that the Exam foci did not embrace all ‘core skills’ (Table 11.3a) and that the Exam was more knowledge-, rather than skill-oriented (Table 11.3b). Consequent to the latter characteristic, student answers in the Exam could sometimes just be learnt by heart, according to the respondents.

As far as Exam administration was concerned, the examiners admitted that students were not always in equal situations, and that cheating ‘sometimes happens’ (Table 11.3e). Cheating can be seen, in some ways, as a result of the Exam format, inasmuch as linguistic questions in Task 1 are available before the Exam and students can prepare for them at home. As those linguistic questions presuppose mostly reproduction of information from coursebooks (Chapter 5), with a very limited element of spontaneity, some students might take steps to produce ‘aide-mémoires’ before the Exam starts. With modern gadgets, this method of cheating may become even easier. Exam administration procedures do not presuppose students leaving their belongings outside the Exam room. Whereas big items like dictionaries and laptops are not allowed, there are no restrictions on smartphones or tablets of smaller sizes. Even in a mute mode, these devices can be used as PDF readers, if the necessary content is downloaded before the Exam. At the same time, cheating is quite unlikely in Tasks 2 and 3, for which students receive texts for reading and listening in the Exam room. With

⁶² Exam threats were singled out as a result of the literature review on language testing (Alderson, 1995; Heaton, 1995; Bachman & Palmer, 2010)

administration procedures described very vaguely, the respondents see the issue of reliability as important. Apart from unequal conditions for students and cheating, examiners and Exam developers question the reliability of examiners – both intra- and inter-rater. With assessment criteria being presented loosely, students performing on different topics and in (sometimes) unequal circumstances, providing continuity of assessment can be seen as a hard task for examiners.

As Table 11.4 illustrates, the majority of respondents see changes in the Examination as desirable – 16 out of 20 people gave ‘yes’ answers to the question ‘Would you like changes to be introduced?’ Changes, according to the respondents, can take place in the following areas:

- Exam content;
- Exam format: excluding some parts, adding new parts, using task types different to those currently employed;
- Exam administration (Chapter 11: Table 11.5).

As can be seen, the Exam developers’ vision of the Final Language Examination does not differ from the vision of other stakeholders (Exam takers and school teachers of English) discussed in Chapter 10. This fact allows for some freedom in suggesting possible changes in the Exam, because they are unlikely to collide with the opinion of major decision makers – namely, Exam developers at the Faculty of Foreign Languages.

The collected data also describes some *contradictions* between the current Final Assessment practices in the studied context and the way the stakeholders see those practices. As Table 11.1 shows, some examiners and exam designers feel satisfied with the current Final Language Exam⁶³ – either with all aspects or with some of them (content or format), although the present data reveal many of the Exam’s weak points. According to some responses (Chapter 11: Tables 11.4, 11.5), the respondents consider some changes as desirable; for example, adding new parts to the Exam or reconsidering the existing parts. So, the discrepancies within the situation under study are found in:

⁶³ At the same time, there are a number of examiners and Exam designers who do not feel happy about the Examination under study

- approximately half of the respondents feeling satisfied with the Exam, although its weaknesses are quite numerous;
- the other half of the respondents feeling that the Exam needs changes in content, format or both, while none of the changes have taken place within at least the past 15 years⁶⁴.

There might be several reasons for the contradictions between facts and attitudes to those facts. Some of these reasons can be seen as social – the stakeholders involved in Exam design and administration are not motivated to make any changes. *Intrinsic motivation* for change might be hindered by some social factors – from constant reduction in the number of students at the Faculty⁶⁵ to quite high work load at each department. Although some inconsistency can be observed between reduction in student numbers and increase in work load, an insight into lecturers' job descriptions can be helpful. Twenty, and even ten years ago, the major part of the work load was teaching: contact hours, supervision of student research and teaching practice. Nowadays, with demographic changes and re-consideration of curricula, the number of teaching hours is decreasing because the number of students is also decreasing, but the amount of paperwork is increasing. Many university teaching staff members find this demotivating. *Extrinsic motivation* in the form of Ministry of Education requirements, university ratings, etc. might be stimulating but, as described in the Introduction, the Ministry provided freedom to universities in terms of defining their training programmes and final assessment, and this freedom allows universities to avoid introducing any changes if university staff do not consider those changes essential.

One more reason for sticking to the current Exam format and content might be considered here, although it needs more evidential support. Some stakeholders cannot see the Exam critically because they have almost nothing to compare it with. Not many of the respondents have access to international experience in language assessment of language teachers, although much information can be found on the Internet nowadays.

⁶⁴ Discussed further in this part

⁶⁵ This can be explained by demographic factors – the number of school leavers decreases every year, and is expected to continue decreasing until at least 2020

The responses to the question about attitudes to the current Final Language Examination might have been different if the respondents had first been offered alternative tasks to compare with the tasks currently in use, and asked to choose those that they considered more relevant/efficient. It must be noted, though, that even if alternative tasks were offered, some respondents might be limited by their beliefs. For example, one very likely reason for the linguistic questions (Task 1) in their current form having been included in the Exam format for at least the past 20 years might be a strong belief on the part of many university teaching staff members that linguistic knowledge is the key component of English teacher development. A general 'belief in knowledge' among Russian teachers can be seen as one of the reasons for the participants in this research favouring linguistic knowledge as a major route to successful teaching. By no means does this thesis aim to undermine the role of linguistic knowledge in FL teacher development. However, it is seen here as part of a prominent discussion in Russia that involves schools, parents, universities, and local education authorities. The majority of participants see 'knowledge-based' teaching as a universal approach to training specialists. It is seen as threatened by competence-based approaches, with the latter being thought of as ineffective and even 'alien to the Russian educational values' (see, for example, <http://argumenti.ru/society/n523/432461>)

The current Examination format and content were introduced at least 20 years ago, when the major foci of teacher development were linguistic knowledge and language skills, with reading playing a dominant role. Speaking was mostly limited to prepared speech (both monologue and dialogue). Integrated reading and speaking tasks very often were tasks on stylistic analysis of literary texts. Courses in Practical Grammar and Practical Phonetics aimed, first of all, at developing accuracy of spoken and written production and development of linguistic awareness, with the latter often becoming the more essential component of these two subjects. Writing was mostly seen as a means, but not as one of the aims, of teacher development; however, the situation changed in the early 2000s when writing became an essential component of the teacher training curriculum. Although this study does not aim to look into the background of the current Final Language Examination and the history of its development, it can be presumed that in the 80s-90s the Exam was a reflection of the programme of teacher development and its aims, as any examination is supposed to

be. With the aims shifting in the 200s towards communicative development of future teachers, the Exam stayed the way it used to be, with only minor changes introduced. For some reason, the gap between new requirements in teacher development, new teaching aids, and some new developments in teacher education and evaluation (e.g. international examinations for language teachers, as presented in Chapter 4) did not make obvious the gap between the current form of Final Assessment at university and the new reality of teachers' functioning in and out of the language classroom.

12.2. Suggestions for possible changes in the Final Language Examination for language teachers (university graduates)

Strengths and weaknesses of the Examination under study (summarized further in Table 12.1) can be divided into two groups:

- *pre-exam issues* like appointment of examiners and Exam designers, and Examination materials moderation and piloting;
- *Exam content and administration*: selecting the content, defining the format; developing assessment criteria; issuing Exam administration guidelines; marking.

The first group of strengths and weaknesses is viewed in this research as containing issues quite difficult to change, should such changes be required. This may be explained by the dependence of this group of factors on documents and regulations as well as by additional financial and time issues. For example, as stated earlier in Chapter 5 and Chapter 7, exam specifications are not considered an essential document for universities in Russia, whereas having an Exam syllabus is obligatory, as prescribed by the State Standards. No universities are expected to provide criteria for the appointment of an examiner/materials designer, so very often it is based on the examiner's experience and appropriate degree, as in the case under consideration. Moderation and piloting of exam tasks are not considered essential parts of the materials design process by any document either at the federal or local level. As a result, any work of this kind would be added to, but not initially included in, lecturers' work-load. Involving other stakeholders, like school teachers, in the process of

materials moderation requires extra money which not many universities are ready to invest.⁶⁶

The other group of issues does not involve extra work for the department staff, but presumes work being done in a different way. For example, instead of selecting 20 texts for reading and 20 audio-texts for listening one set of tasks can be designed which would be based on 2-3 texts with tasks assessing different reading strategies.

Suggestions for some changes in the Exam were informed, on the one hand, by the factual data and opinions of various stakeholders: examiners, Exam developers and administrators; Exam-takers; and teachers of English with different amounts of experience (from 1 to 40 years in language education). On the other hand, the data obtained were weighed against national and international experience in language teacher assessment (Chapter 4). This resulted in a number of suggestions for change in:

- examination *content*
- examination *format*
- exam *administration*
- procedure for exam task design.

Changes in the Examination **content** may be treated as those that are quite easy to implement from the administrative point of view. The reason for this statement can be found in the State Educational Standards (2010) – the major educational document issued by the Ministry of Education of Russia:

'8.6 Final assessment should include submission and presentation of Graduate Thesis (Выпускная квалификационная работа)⁶⁷. Final examinations are administered at university's discretion. Format and content of the Final examinations are defined by university' (2010: 13)

As can be seen, universities have some freedoms to define content and format of Final examinations in general and language examinations in particular which immediately leads to freedoms for exam designers: no restrictions can be traced on the choice of task types, texts and other input for exam questions, or on expected outcomes.

⁶⁶ Writing Exam specifications which would include criteria for task and text selection, description of expected performance and all other issues expected in test specifications, administering materials moderation and piloting can become an independent project on exam materials design for Russian universities (FL departments)

⁶⁷ The graduate thesis is written and presented in Russian. Marks that graduates get for their theses do not influence the marks on any other Final Examination(s)

No indication of any restrictions on the Exam content has been found in another document – the Final Examination Syllabus designed by the FL department. Therefore, in order to meet one of the major points of criticism of the Exam content – absence of professional component and misrepresentation of communicative skills under assessment together with unnecessary complexity of the linguistic part – changes can be introduced at the level of the FL department with no further agreement with the LEA or Ministry required.

Thus, the changes in Exam content might take place through:

- reconsidering and widening the range of texts for listening and reading
- reconsidering task output, i.e. instead of pure retelling of the content, a task might involve relating the issue discussed to a concrete situation OR reading 2 short articles with different views on the same issue and comparing/contrasting them;
- widening the range of skills and tasks⁶⁸ and including more teacher-related items (e.g. finding a coursebook in a catalogue to suit the situation provided by the Exam task; correcting errors in a piece of written work; evaluating difficulty of a text/task);
- reconsidering the current extensive linguistic questions and/or adding teacher-specific language awareness tasks.

Changes in the Exam content could involve reconsideration of task design procedure (see further in this chapter); and, in particular, would require a more detailed description of task and text selection criteria. All these would cause changes in the Final Language Examination syllabus and might include design of:

- a list of topics to be covered at the Examination and a balance between ‘general’ and ‘teacher’ English.
- criteria for text selection and their weight
- sample exam tasks and description of expected outcomes
- recommendations for task designers (number of tasks for each section, balance of skills, etc.)

Changes in the Exam **format**, being closely connected and intertwined with Exam content, have been touched upon previously in terms of widening a possible range of task types and

⁶⁸ The suggestion refers to the format of the Examination but is discussed here for convenience and follow the logic of the discussion

including more teacher-oriented tasks. Nevertheless, there is at least one more issue to be discussed – possible changes in the form of the Exam and its transition to ‘written and oral’ instead of ‘oral’. Although State Educational Standards give universities freedom in defining exam content and format (2010:13), this freedom can be treated as an opportunity to leave everything the way it used to be for many years. Therefore, introduction of a written part, which is a serious issue by itself, might cause some problems:

- introduction of a ‘Writing’ part with an essay or other written assignment would cause changes in the duration of the Exam and changes in the Exam administration; in the design of assessment criteria and marking scheme; and in means of striking right balance between the oral and written part;
- employing a wider range of task types, including close-ended ones for listening/reading and, possibly, for the linguistic part (Chapter 4), would result in re-consideration of the whole marking framework which would then include both keys for close-ended tasks and assessment criteria for open-ended ones.

A possible issue that might arise if the Exam format becomes mixed and a written part is added is *marking*. As stated previously in Chapter 5, marking and grading are currently done and the marks announced on the day of the Final Examination. Marking students’ written answers will definitely take extra time, besides which, time will be required for marking oral answers and grading each student. These would either require a bigger team of examiners (to mark written papers while other members of the team listen to oral answers) or announcement of marks within the next few days. Such changes might be viewed as challenging for various reasons:

- it will require re-planning of examiners’ work load;
- some training of examiners in how to mark written work might be required, especially if a revised set of criteria is introduced;
- the suggested new format differs significantly from current Exam practice, so stakeholders (examiners and administrators) might require time to realize why the changes are taking place and that they are essential, and why.

Changes in Exam format would lead to changes in Exam **administration**. Thus, introduction of a written part would immediately require a set of criteria, samples of student writing for piloting these criteria, redesigning the time frame for the exam, and re-distribution of roles within the examination team.

Irrespectively of whether or not a written part is introduced, a set of criteria for assessment of students' oral performance would be essential. Even with minimal changes in speaking tasks, reconsideration followed by clarification of assessment criteria is seen by many stakeholders as a very important step in Exam improvement. If the changes in speaking tasks are more profound, a new set of criteria will also be of key importance.

Although it is not always treated as an aspect of exam administration, training of the Exam team is seen as a very important step in Exam improvement. It is necessary not only for examiners who apply assessment criteria to student answers, but also for Exam designers, because analysis of the current situation demonstrated lack of clarity in task types and text selection, with rather vague understanding of moderation and piloting.

Possible changes in the Exam are summarized in Table 12.1.

Table 12.1: Possible ways of changing the current Final Language Examination

Current situation	Possible changes (based on analysis of empirical data)
Content and format of the Examination	
Overly complicated linguistic questions (Task 1); this amount of knowledge is not required at school	Linguistic questions reconsidered /reformulated, for example: <ul style="list-style-type: none"> • by being made less extensive (requiring less information to be reproduced) • by being presented in another form (e.g. multiple choice; gap-filling tasks for testing terminology and linguistic awareness; tasks based on reading an extract of an article on linguistic issues) • by adding teacher language awareness tasks instead of 'traditional' questions (e.g. correcting errors in samples of student performance)
Limited number of language skills assessed	Wider range of skills introduced with a wider range of tasks, for example: <ul style="list-style-type: none"> • in addition to reading and listening for detailed understanding, gist and specific information tasks included • an unprepared monologue task or a prepared lesson fragment introduced; • vocabulary tasks, including teacher vocabulary, introduced (multiple choice, matching, gap-filling, giving definitions, etc.) • writing skills included (a lesson plan; an official letter; a summary of an ELT article; note taking or an essay)
Range of texts not wide enough No coverage of professional component	Texts other than newspaper articles employed for assessment of reading, for example: <ul style="list-style-type: none"> • articles from ELT/linguistic journals, including those online; and from non-professional magazines • extracts from teachers' books; reference materials;

Current situation	Possible changes (based on analysis of empirical data)
	<ul style="list-style-type: none"> • samples of student writing for giving feedback; • fiction <p>Range of listening texts widened, including samples of student performance</p> <p>Text and task selection criteria are made clear</p>
Exam administration and marking	
No administration guidelines on timing, examiner behaviour	<ul style="list-style-type: none"> • Separate sets of criteria for assessing linguistic knowledge and language skills (for the format similar to the current one) • Keys to objective tasks (for example, reading; listening; vocabulary and terminology if introduced in a new format)
No set of assessment criteria; no descriptors for criteria	<ul style="list-style-type: none"> • Sets of criteria for assessing speaking and writing (if there is a writing part) • Descriptors for criteria – ‘can do’ or ‘expected performance’ type
No commonly accepted ways of resolving disagreement	<ul style="list-style-type: none"> • Weight of criteria and tasks clearly defined before the Exam • Examiner behaviour specified (possibility of intervention; kinds of intervention)

As can be seen, changes are considered necessary in practically every area of the Final Language Examination – from development of tasks to administration and marking. However, not all these changes can be implemented simultaneously, due to various factors. In addition, some changes are easier to implement than others. Further in this thesis the reasons for this are discussed and changes are considered as essential/desirable/optional.

Chapter 13

Suggestions for possible changes in the current Final Language Examination for future language teachers

This chapter summarizes the strengths and weaknesses of the current Final Language Examination and maps out several forms of its possible change. The chapter considers 2 groups of changes – minor and major - that would require different amounts of time, resources and people involved. Chapter 13 also presents some alternative Exam tasks and discusses their relevance and applicability for the language examination to language teachers.

The chapter consists of 3 parts. First, suggestions for possible changes are considered and prioritized, with examples of alternative assessment tasks provided. Then, possible outcomes of such changes are analysed – from changes in the Examination itself and their effect on the Exam’s validity, reliability, authenticity and practicality, to changes in attitudes towards the Examination in particular and language assessment practices in general. The chapter concludes with a discussion of the limitations of this research.

13.1. Possible changes in the Examination content, format and administration

As demonstrated by the empirical data, the current Final Language Examination requires changes in several areas – from its design to ways of marking student answers. Chapter 12 (p.239) suggested at least 2 dimensions that needed reconsideration:

- Exam design issues, including choice of format;
- Exam content and administration.

It can be presumed that one group of changes would automatically involve the other. Thus, changes in the Exam format might influence Exam administration procedures. Reconsideration of the Exam foci might require a wider range of tasks, and assessment criteria other than those currently used. Table 12.1 in Chapter 12 (pp.243-244) presents

all possible changes in the Exam, although it is understood that not all of them can be put into practice immediately.

In this research, several directions for changes are mapped out and considered:

1. Minor/ restricted changes.
2. Major/ more global changes.

The *first* group involves changes in the Exam design and, partially, its administration, in terms of specifying assessment criteria. The *second* group comprises changes in the Examination content and format that might lead to complete or major reconsideration of its administration and the whole assessment procedure.

The first group of changes seems more feasible and achievable within a limited period of time (e.g. one academic year). It could start with the design of detailed Exam specifications which, in addition to the existing Exam syllabus, would include:

- knowledge areas and skills under assessment and their representation through the Exam's content and format;
- principles of input selection – choice of texts and tasks, requirements for texts – source, length, genre, topic/issues discussed, etc.
- representation of topics and balance between general and teacher English;
- assessment criteria/ keys for each part of the Exam; weight of each criterion;
- Exam design procedures, including task moderation and piloting;
- examiner behavior – giving clues to students, asking additional questions outwith the Exam tasks, etc.

The design of Exam specifications might require training for the Department staff, especially on issues dealing with Exam administration, examiner behavior and Exam design.

The existing form of the Final Language Examination would not allow several issues to be resolved. One such issue is timing – as regards both the amount of time for students to prepare their answers and the time they spend answering the Exam questions. Specifying examiner behavior in the ways suggested above might clarify Exam administration procedure. For example, if the number of additional questions is standardized or such questions are not recommended at all, the time that students spend answering is likely to fit the time frames selected.

Design of assessment criteria could be another issue to be reconsidered. With the existing Exam format, such reconsideration can start with a clear division between assessment criteria for Task 1 (Linguistic knowledge) and Tasks 2 and 3 (Listening and Reading). Design of assessment criteria for the Linguistic knowledge task would involve parameters usually considered within Content and Language Integrated Learning (discussed in Chapter 4, pp.94-96). The issues to consider here are similar to those singled out in previous research (e.g. Gablasova, 2014). If assessment is administered in a FL, how is it possible to determine whether it is lack of knowledge or of language skills that prevents Exam takers from demonstrating the knowledge? What criteria should have more weight: those referring to the content or the form? If the form (e.g. Speaking skills) is important, what parameters need to be assessed? – accuracy, fluency, range of language means, etc. These questions are not to be answered in this research, as they involve decisions taken by the Department of Foreign Languages. Once such a decision is taken, it would be possible to design and pilot a set of assessment criteria for Task 1.

Task 2 and Task 3 allow for design of similar sets of assessment criteria that could include:

- degree of understanding of the input texts (listening/reading);
- the content and clarity of the summary for the input text;
- accuracy and fluency of student oral performance;
- range of linguistic means employed.

Table 13.1 presents an attempt to design assessment criteria for Tasks 2 and 3 in their existing form (Chapter 5). The table aims to show one of the ways assessment criteria could look, and is not considered as a final version to be employed by the examiners.

Table 13.1: Assessment criteria for the current Final Language Examination (Task 2: Reading, Task 3: Listening)

Reading/listening					
	5 (excellent)	4 (good)	3 (satisfactory)	2 (poor)	
Understanding of the text	Full understanding of the text and ideas expressed by the author	Good understanding of the text, although Exam taker has ignored some details	General understanding of the text content; some facts/author's ideas are misinterpreted/ misunderstood	Superficial understanding of the text; Exam taker misinterprets the key information in the text	
TOTAL for text understanding:					
Speaking					
	5 (excellent)	4 (good)	3 (satisfactory)	2 (poor)	
Content	Exam taker presents the content logically, with significant amount of detail. No irrelevant information is presented. Exam taker presents the content of the text and his/her vision of the issues presented there	Exam taker presents the content logically, although some details are omitted. Exam taker doesn't discuss in detail the issues presented in the text	Exam taker retells the content of the text without discussing its major issues/ without expressing his/her attitude to the issues presented by the author	Exam taker cannot retell the content of the text. No vision/ interpretation of the issues is presented	
Range	Lexical	Wide range of vocabulary, including idiomatic expression, that results in colourful speech and ability to communicate ideas freely	Sufficient range of vocabulary; Exam taker expresses him/herself with ease, although there might be some repetitions	Quite limited lexical range but Exam taker conveys his/her ideas	Very limited vocabulary that causes difficulty in expressing thoughts and ideas
	Grammar	Wide range of grammar structures that allows Exam taker to express ideas freely	Sufficient range of grammar structures	Quite limited grammar range. Exam taker mostly operates at a simple sentence level	Exam taker sticks to 2-3 grammar structures and finds it difficult to express his/her thoughts
Accuracy	Phonological	Good articulation of sounds and proper use of intonation patterns. 1-2 slips that do not interfere with understanding	3-4 repetitive mistakes in pronunciation of sounds. Mistakes do not interfere with understanding	Frequent mistakes in pronunciation, which do not lead to misunderstanding OR 1-3 mistakes that lead to listeners' confusion	More than 8 mistakes in pronunciation that may lead to misunderstanding
	Lexical	Words/ lexical units including idioms are used correctly	1-2 lexical mistakes that do not lead to misunderstanding	3-4 lexical mistakes or 1-2 repetitive mistakes that lead to misunderstanding	5-6 lexical mistakes that lead to serious confusion

	Grammar	No grammar mistakes or 1-2 slips	1-2 repetitive grammar mistakes	3-4 repetitive mistakes	5 and more grammar mistakes
Fluency		Natural tempo with natural pauses. Exam taker expresses thoughts and ideas freely	Natural tempo; some hesitation pauses that do not lead to any communication breakdown.	Frequent hesitation pauses or slow tempo but Exam taker manages to express his/her thoughts and ideas without major strain on the listener	Frequent pauses caused by search for language means to express thoughts and ideas. Slow tempo and/or many unfinished sentences that cause serious strain on the listener
TOTAL for speaking:					
TOTAL:					

Similarly to the need to provide training for Exam designers, design and use of assessment criteria may also require some consolidation. This statement is based on the experience of implementing the National Examination in schools, which disclosed considerable difficulties for teachers in using the designed set of assessment criteria for the Speaking and Writing parts of the National Exam on Foreign Languages.

The second group of changes – major changes – would start with reconsideration of the content of the Final Language Examination. As suggested by the stakeholders, and school teachers of English in particular (Chapters 9, 10), changes in the Exam content might include a wide range of options:

- omitting the linguistic part *OR* its reconsideration in terms of simplifying the tasks and adding the professional (teaching) dimension;
- introducing Teacher Language Awareness tasks that may replace the ‘traditional’ linguistic task;
- widening the range of tasks for the Listening and Reading sections;
- separating Listening and Reading tasks from Speaking tasks, although such a separation might in some respects be considered artificial from the communicative point of view;
- introducing tasks for testing professional vocabulary and Classroom English;
- introducing productive writing tasks – from a formal letter/email to an essay.

To illustrate each of the above suggestions, alternative Exam tasks were developed. Similarly to the alternative assessment criteria presented in Table 13.1, the alternative tasks are seen as a springboard for further development of the Examination, and require the opinions of Exam designers, examiners, potential Exam takers and practising teachers of English to be obtained. Design of the alternative tasks for the linguistic part of the Examination was informed, on the one hand, by the opinions of the stakeholders and, on the other hand, by the outcomes of the language examination review (Chapter 4).

Alternative task 1 can be seen as a revised version of the existing linguistic task and will require a minimal number of changes in the Examination. Thus, instead of a complex question on one of the linguistic issues (see the example below) it is presented in a slightly different form that would require not only linguistic knowledge but also Teacher Language Awareness.

#1 Task type – an open-ended question

<i>Current task</i>	<i>A possible option</i>
Speech sounds. Classification of vowels and consonants. Principles of classification	English sounds: vowels and consonants. What difficulties might Russian learners of English face when mastering English sounds? Which sounds can be more difficult to teach? Why? How can you deal with these difficulties? Give two examples.

Alternative task 2 was informed by similar tasks in some international examinations for language teachers (e.g. Praxis®, ICELT, TKT) and aims to assess Teacher Language Awareness (TLA) plus the ability to spot errors and provide feedback on student written performance. The task is based on authentic student output demonstrated by a pupil at one of Tula’s comprehensive schools.

#2 Task type – error correction

**Read the composition written by a 9th former (15 years old, secondary school).
On your Answer Sheet, mark the mistakes and define their type:**

- Sp – spelling
- Gr – grammar
- Punc – punctuation
- WW – wrong word

Are there any errors that lead to communication breakdown? What are they?

My summer spent very funny. I worked and travelled a long time. First month I spent for cottage helping aunt with repair. I read fantastic and classic literature.

In many wrote in July. First part of month I got involved draw and often drew a night. I drew still life, nature and person from comics. Second part of July I was in home relax and spent fine time. I many played in volleyball and swam in the river. Many friends appeared with which I still communicate. I went on fishing. Every evening we watched films or go on disco after dinner. In rainy weather we stayed in house and could watch TV, played table games or drank tea and told each other funny stories.

In August I rested in south, in Crimea. Brother told me a lot about his studies. We were on beach or swam in the Black Sea for a long time. We walked to the port or went to concerts every evening. About my summer I have a lot of impressions.

Alternative task 3, also based on an authentic student exercise, was informed by the TLA tasks in Praxis and TKT examinations for teachers.

#3 In each sentence, identify the type of error:

1. My holidays began and it's not necessary to wake up early and go to school.
 - a) grammar (including use of prepositions)
 - b) punctuation
 - c) spelling
 - d) wrong word
 - e) other
2. This summer I got a job and worked two weeks.
 - a) grammar (including use of prepositions)
 - b) punctuation
 - c) spelling
 - d) wrong word
 - e) other
3. Also in the village there are neighbours' pets: chicken, ducks, cows and sheep.
 - a) grammar (including use of prepositions)
 - b) punctuation
 - c) spelling
 - d) wrong word
 - e) other
4. Every day we with friends was at my house and drank tea with sweets.
 - a) grammar (including use of prepositions)
 - b) punctuation
 - c) spelling
 - d) wrong word
 - e) other

These tasks on linguistic awareness and teacher language awareness are seen as a response to the stakeholders' feedback on the current Task 1 of the Final Language Examination. A test on linguistic concepts was offered by one of the interviewees as an alternative task type (Chapter 9) but no samples of such tests are presented in this research. The design of such a test is seen here as a task requiring a joint effort by TESOL and Linguistic teams.

Another section of the Examination that might undergo some considerable changes comprises its Reading and Listening parts. The easiest option seems to be widening the range of texts and including 'teacher-related' texts that would be both relevant and interesting for future teachers of English. Two samples of the Reading task are presented in Appendix 15 (*Alternative tasks 4 and 5*, pp.358-361). As can be seen, the input texts were reconsidered, whilst the task type remained the same as in the current Reading part of the Exam.

Bearing in mind that not only input texts but task types as well were criticized by some respondents, alternative tasks were developed to involve a wider range of skills and reading strategies within both general and professional English. *Alternative task 6* (Appendix 15, pp.362-363) serves as an example of a Reading task that involves reading an ELT journal and spotting specific articles in the table of contents. According to the results of the Needs Analysis, reading ELT literature is one of the activities that teachers deal with regularly.

A similar task of a lower level of difficulty (*Alternative task 7*) is presented on p.364) of Appendix 15. This task aims to test reading for general understanding by offering a set of items to match within the ELT area.

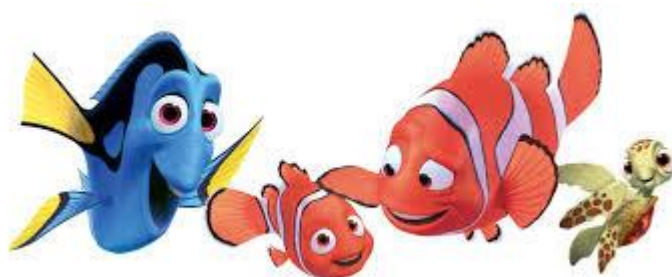
Reading for detailed understanding can be assessed through *Alternative tasks 8 and 9* (Appendix 15, pp.365-371). Task 8 offers an extract from the Teacher's Book for one of the most commonly used coursebooks designed by a team of Russian and British authors – New Millennium English.

Appendix 15 presents more alternative tasks for assessing reading. As mentioned earlier in this chapter, there is always concern about dividing skills into listening, reading, speaking and writing and testing them as separate skills, whereas in real life

we mostly deal with integrated tasks. Another concern refers to language task design in general and content selection in particular. This concern is widely discussed in papers on designing TOEFL tasks: how to introduce some professional dimension into reading and listening texts without making them professionally biased and undermining their validity and reliability (e.g. Brown, 2000). The same issue is addressed by the Content and Language Integrated Learning (e.g. Coyle, 2010; Gablasova, 2014): how to test language skills without the knowledge of content interfering? TOEFL seems to have found a solution by using texts that are not heavily loaded with terminology and do not require specific knowledge in order to understand them (e.g. Alavi&Akbarian, 2012; DeLuca&Cheng, 2013).

Bearing these concerns in mind, some alternative tasks were designed to integrate reading and speaking skills, as presented below.

#10 Below is an article about using technology in teaching English to young learners. Read it and summarize the author's main points. Do you agree with the author? Why/why not? What, do you think, is the role of technology in a language classroom?



Finding Nemo on the iPad

Michael Tasseron
michael.tasseron@gmail.com

As learners increasingly make use of smartphones and tablets outside of the classroom, it should be expected that teachers also adopt such technologies for use in their lessons. One way the iPad can be used is through storytelling, and a wide range of colourful animated and interactive e-books are now available for this device which will bring stories to life in the classroom. An e-book application which I use regularly in my lessons is Disney's *Finding Nemo*. It can be used successfully with learners aged four to seven. The approach I use for this is guided by the recommendations made by James Bourke in an article in *ELT Journal*, where he argues the case for a topic-based syllabus and the use of materials which should relate to the world of the young learner, where '*there are no tenses, nouns or adjectives*'.

The value of stories

Stories have immense educational value, and James Bourke recommends that they should be regarded as an essential element in a young learner syllabus. I am also of the opinion that they should be a regular feature of lessons wherever possible, even where the syllabus does not make provision for them. Children love stories and, when they are used effectively, they are bound to engage young learners of all ages.

Procedure and presentation

Fortunately, most children will be familiar with the antics of Nemo, the lovable little clownfish, whose desire for adventure lands him in trouble and who finds himself in a fish tank many miles from home. His father, Marlin, then embarks on a long and perilous search for him, which takes him all the way to a dentist's office in Sydney.

I try to leave ten to 15 minutes at the end of my 50-minute young learner lessons for short stories. My typical class size ranges from four to six learners, but this type of storytelling can also be used for slightly bigger groups.

The Finding Nemo e-book is interactive and has music and a range of sound effects which accompany the story. When you touch one of the characters on the iPad screen, the name of the character is played. The music used also corresponds to the tone of each scene, and conveys emotions such as excitement, relief or sadness.

Accompaniment and appropriacy

A useful addition to using the e-book app is a picture dictionary which includes a page about sea creatures, although this is not essential. Prior to starting the story, I open the picture dictionary to the sea creature page and ask the learners what they can see. The aim of using a picture dictionary is to familiarise the learners with the story content. The teacher can also ask questions about the colours and sizes of the different creatures. If a picture dictionary is not available, similar questions can be asked about the host of colourful characters who appear throughout the book. With regard to the learners' responses, the emphasis should be on comprehension. Therefore, short one-word answers are sufficient. Questions such as *Is the shark big or small?* facilitate such responses.

The iPad is a wonderful tool, and in the appropriate setting learners can be encouraged to interact with it. However, caution is advised. Allowing learners to touch the screen to turn the page or find out a character's name may seem like a great idea, but it is best left to the teacher to do this. I find it more efficacious to ask the learners what they hear when I touch the screen and a character's name is played, or what they can see on the screen. The reason for this is that learners in this age group are easily distracted by different sources of stimulation. They are also not usually able to remain focused on what the teacher is saying, while at the same time using the iPad. Also, bear in mind that their young brains are trying to process information being presented to them in a foreign language. Thus, distractions caused by squabbles about who can touch a colourful character can disrupt the chain of thought and result in a loss of interest or discontent. Stories provide opportunities for children to dream and enter into a world outside of direct teaching and learning. As such, it should be a time where everyone can be comfortable and enjoy the magic.

Repetition and revision

While reading, it is a good idea to repeat the names of the characters and what is taking place, to help the learners to follow the story. When you have finished the story, the learners are likely to request that you read it again in subsequent weeks. If you are using a number of

different e-books, you can ask the learners which story they want to read. This is again another opportunity for authentic communication in the form of negotiation. Furthermore, repeating a story is advantageous as it serves to review the linguistic content the learners are exposed to.

★★★

As teachers, we all know that our learners enjoy variation, and making use of new technologies such as the iPad, combined with traditional teaching approaches, can work wonders in engaging our learners. Storytelling is one example of this, and literally any story can be transformed into an exciting part of the lesson, with animated characters that move around, accompanied by sound and music. Young learners are bound to relish this. Apart from the aspect of enjoyment, such stories are also a valuable learning tool, as they expose learners to authentic language use and allow for both direct and indirect learning opportunities and negotiation, as well as revision.

• www.etprofesional.com • ENGLISH TEACHING *professional* • Issue 87 July 2013

As suggested by some respondents, particularly those who had graduated from the College–University programme (Chapter 1), some alternative speaking tasks were developed. They were informed by some international experience in speaking task design (Chapter 4), the existing international language examinations (e.g. Cambridge ESOL) and some experience gained through the Tula PRESET project run by the British Council (1998-2007).

The tasks below are based on different types of input (verbal and/or visual) and various levels of expected performance – from short answers to interview questions to extended responses to problem situations within both general and teacher English.

#11 Interview based on verbal input

1. What age group would you prefer to work with? Why?
2. What do you think you can do in and outside the classroom to provide effective learning?
3. What would you try to avoid in your work?
4. What opportunities do language learners have these days? Are they different to those you had when you were at school?
5. Do you think you'd be a good teacher? What makes you think so?

#12 Interview based on visual and verbal input

1. What countries do you think are presented in the pictures? Do you think the classrooms you see are typical of these countries?
2. What similarities and differences can you think of when you look at these pictures?

3. Do you think these are mono- or multilingual classes? What advantages of teaching mono-/multilingual classes can you think of?
4. What resources would you mostly rely on if you taught these classes? Why?
5. What activities, do you think, would work well? Why?
6. Are there any activities you wouldn't use with these classes? Why?
7. What class(es) would make you feel more comfortable? Why?



#13 Monologue based on verbal input

You are preparing a video for your colleagues in other countries to present on social networks. On the video you want to show a selection of coursebooks or other teaching resources that are widely used for teaching English in Russia. Discuss what you want to include in the video.



Your friends are thinking of sending their primary school child to a language school. They asked for your advice, as there are at least 5 language schools in your city, and the parents are a bit lost. They don't mind your contacting these schools for them. What information do you need to give the parents sensible advice?

The set of tasks that are completely new for the current Final Language Examination comprises *writing tasks*. *Alternative task 14* presents an example of an integrated task requiring TLA and professional writing skills. This task can be seen as a development of Alternative task 2 – in addition to the ability to spot errors, Task 14 assesses the ability to provide adequate written feedback.

#14 Read the composition below (written by a 15-year old secondary school student) and identify the mistakes. Write a paragraph, giving the student advice on what needs to be improved.

Every summer I rest my grandmother in the country. There are a lot of my friends. They come on holidays as I am. We go to the forest for mushrooms and berries, swim in the river. This summer, the guys taught me how to play football. I became a very good player. Once we found in the garden a little kitten. I like animals very much. So I took the kitten in the house. He drank the warm milk, slept and mew. The boys went to the river. When they returned the kitten was not at home. We long searched everywhere. Suddenly Andrew heard something. It was a kitten, who climbed on the roof. I took a high ladder neighbours', and climbed to the roof. The kitten was trembling from fear. I took him in hand and we slowly came down. My little friend was afraid. We fed him with milk.

notes

More alternative assessment tasks are presented in Appendix 16. The alternative tasks would require moderation and piloting, as well as obtaining feedback from practising

teachers. If the tasks prove successful, they would require the design of assessment criteria and keys to the close-ended tasks (e.g. Reading).

The changes above might involve changes in the form of the Final Language Examination and turn it from an oral into a mixed one – oral and written. Whatever changes are introduced, they would influence the system of assessment criteria and Exam administration. The key challenge here is seen as agreement between potential Final Assessment tasks and the tasks employed in teaching and progress assessment. If such an agreement is not achieved, the content validity of alternative Exam tasks can be questioned.

13.2. Implications for changes in the Final Language Examination

As can be seen from the empirical data obtained (Chapters 7, 9 and 11), documents analysed (Chapter 5) and the current Exam as described and compared to national and international experience in language teacher assessment, a list of suggestions for the Final Language Examination is substantial and deals with practically every dimension of the Examination – from its design to the announcement of final marks. The issue to be discussed in this part deals with the importance of these changes and the longer-term outcomes that these changes might result in.

Although this research only deals with one exam – the Final Language Examination – at the Faculty of Foreign Languages, the impact it currently has and, therefore, the impact that changes might bring is much wider. For convenience and logic, possible outcomes of Exam improvement are discussed here in 2 groups, but this is done for research purposes only, and in reality the effects of the two groups can intertwine and depend on each other. *The first group of outcomes* deals with the Examination itself and its key characteristics – validity, reliability, authenticity of tasks and practicality. *The second group of outcomes*, which might be called ‘attitudinal outcomes’, contains those which mostly relate to stakeholders’ perceptions of what is going on at the Final Language Examination, what the Exam assesses and how it prepares students for future professional life.

The first group of outcomes refers directly to the Final Examination and improvement of its major parameters. As discussed in Chapter 4, validity, reliability, practicality and authenticity are considered key characteristics of any test/examination. Empirical data presented in Chapters 7, 9 and 11, and discussion of the current situation (Chapters 8, 10 and 12), demonstrated that many dimensions of the Examination need to be strengthened.

Validity of language examinations is quite a contradictory issue, according to several researchers (e.g. H.D.Brown, 2004; Norris, 2009). The current Final Language Examination, with its strong and weak points, cannot be considered an exception. As far as *content validity* of the Exam under study is concerned, agreement has been traced between the Final Exam Syllabus and syllabi of theoretical subjects: Theoretical Phonetics, Lexicology, Theoretical Grammar, History of English, and Practical Course of English. In other words, the Final Language Examination does assess what, on the one hand, it claims to be assessing and, on the other hand, what is taught throughout the course of studies. However, according to the State Standards, ‘conditions for assessment should be as close as possible to [student] professional functioning’ (2010: 13), i.e. the Final Language Examination should be a language examination for teachers but not an examination in general English. This statement might mean that, although the content of Final Exam Syllabus agrees with the content of the course of studies, they both lack some professional component.

Closely connected to and, in some way, inseparable from content validity is the issue of **construct validity**, i.e. ‘the relationship between theoretical models and operational assessment frameworks’ (Chalhoub-Deville, 1997: 3) or, after J.D.Brown (2000), ‘demonstrating that a test is measuring the construct it claims to be measuring’ (2000: 8). As Chalhoub-Deville emphasized, there is no need ‘to pigeonhole the components of these assessment frameworks into... a model, but it is expected that the components (...) should concur with the theoretical model that best represents the field’s current state of knowledge’ (1997: 13). Review of publications in the area of language teacher language competence (Chapter 3) indicated that there were several views on language teacher language competence, and no consensus reached on its structure and components. The working definition of language teacher language competence employed in this research includes the following elements:

- teacher language awareness (Wright&Bolitho, 1993; Widdowson, 2002)
- communicative skills of listening, reading, speaking, writing (Thomas, 1987, Elder, 2001; Sešek, 2007; Consolo, 2008; Richards, 2010)
- Classroom English (Sešek, 2007; Richards, 2010).

Another set of data that adds to concern about the construct validity of the current Final Language Examination comprises empirical data from the Needs analysis. It demonstrated rather a vague relationship between what teachers require in and out of the classroom and what they are expected to perform at the Exam. It is hoped that possible reconsideration of the Final Language Examination in terms of assessment foci and content could contribute to the validity and efficiency of final language assessment at the Faculty of Foreign Languages. In the long run, changes in the final language assessment might cause changes in progress assessment and if necessary, in the whole FL teacher training programme.

The issue of **authenticity** of the Exam tasks in some way goes together with validity and for this research it means the degree to which exam tasks are related to the FL teacher's job and the extent to which exam tasks resemble the activities teachers are expected to perform. Similarly to the previously discussed Final Exam's validity, the empirical data obtained through the Needs Analysis cast some light on the authenticity of exam tasks by demonstrating little connection between the tasks graduates perform in the Exam room and in the real language classroom. Although it is fully understandable that creating, in a university exam room, an environment that resembles a language classroom is a next to impossible task, there are still some ways of improving the current situation:

- introducing teacher language awareness tasks;
- using teacher-related texts for reading/listening and revisiting expected outcomes of tasks, for example employing task types other than retelling of what was read/heard⁶⁹;

⁶⁹ Referred to by teachers at interviews as a task they never or very rarely perform

- reconsidering Speaking tasks which, instead of retelling, would involve additional skills with a different expected output (e.g. commenting on a choice of a classroom activity; explaining the choice of a coursebook; comparing/contrasting 2 classroom situations).

Such an attempt was made in this research by developing alternative Examination tasks (part 13.1).

Another dimension of the Final Language Examination discussed in this section is its **reliability**. It is an issue that received quite a substantial amount of criticism, as can be seen from the empirical data (part 9.2). The major concerns were

- absence of a transparent system of assessment criteria;
- unregulated, and sometimes unobserved, timing issues;
- cheating (mostly for Task 1);
- quite low inter-rater and intra-rater reliability for various reasons, mostly the absence of a commonly accepted assessment scheme;
- unspecified examiner behaviour.

Investment in Exam reliability will involve, first of all, quite substantial changes in Exam *administration* (design of assessment criteria, scoring system, keys), alongside possible reconsideration of Exam format. Another issue to be tackled is *examiner training* or at least establishing more effective co-ordination and co-operation within the Exam team in both Exam design and exam administration. Organizing training for the department staff might require extra effort but would result not only in higher Exam reliability but also in its validity. Improvement of these areas, along with other changes described above would be beneficial for the quality of the Exam and, in the end, for different groups of stakeholders, from Exam designers to Exam takers.

The other group of outcomes, ‘attitudinal’ outcomes, might seem to be a particularly far-fetched type of impact that changes in the Final Examination might have on teacher development. Pre-service teacher training in Russia, like training of other specialists, is a transitional stage between school and professional life⁷⁰, when, just after finishing school, teenagers (usually at age 16-18) choose their profession. With big changes in the system of school education (new coursebooks, technologies, approaches to

⁷⁰ For reasons discussed earlier (Introduction, Chapter 2): at the average age of 17, when entering a university, applicants choose both their future degree and their teacher qualification

teaching) and changes in school licensing and accreditation in general and introduction of the National Examination for core subjects in particular (see Chapter 2; Appendix 2), the current assessment format and the whole course of studies at university is sometimes perceived by new students as a minor step back. At the age of 16-17 school leavers take an English examination which had both written and oral parts with all 4 skills – listening, reading, speaking and writing – being assessed; assessment criteria being available and timing issues defined; examiners being external and the exam itself taking place in an examination centre. At the age of 20-21 university graduates take a Final Examination which assesses a great deal, but with teacher-unrelated reproductive knowledge in Task 1 prevailing; with speaking tasks being totally prepared; and with no written tasks at all. The discrepancy might lead to seeing the Exam as an unavoidable step in gaining the qualification of an English teacher, with the assessment focus having no links to the teacher's job. This statement is based on anecdotal evidence only and is not intended as a generalization or as unproductive criticism. However, this dimension might become a separate field of research on school leavers' academic perceptions during the first year of university in Russia.

Another outcome, to take a look ahead, is university graduates' future careers, as many of them seek another degree abroad (mostly in Europe) or are required to take an international language examination (very often one of the Cambridge ESOL exams) by their employers. This is where many graduates experience a big gap between what they were supposed to do (and did successfully) at university and what they are expected to demonstrate at a university overseas or in an international examination room. A counter-argument to this could be the fact that not all graduates go abroad for further study, and assessment systems at a Russian university are not supposed to follow the guidelines of the Cambridge ESOL or some other examination body. There is much reason in this argument, but with Russia joining the Bologna declaration (2005) and, therefore, agreeing to some standards and regulations on the one hand, and, on the other hand, needing to assess much more than reproductive knowledge, some changes in the Final Language Examination for future teachers seem essential.

Table 13.2 summarizes the suggested changes in the Examination by comparing the current situation with a more desirable one, and specifies outcomes that may result from these changes.

Table13.2: Possible effects of changes in the Final Language Examination for language teachers (university graduates)

	Current situation	What it might become	Outcomes
Content	Limited number of skills tested No, or very limited, professional component Extensive linguistic task with no relation to teacher's job	A wider range of communicative skills tested: listening, reading, speaking, writing A wider range of topics employed Teacher Language Awareness tasks (◀ Chapters 3, 4) instead of purely linguistic tasks	Wider range of skills in the focus of attention at the Exam and in the course of studies Students' better understanding of what is expected of teachers at school 'Real life' teacher tasks preparing graduates for the first career steps
Format	A very narrow range of tasks employed Inconsistency of forms: an oral examination presupposes written preparation of answers 2 out of 3 tasks test very similar skills	A wider range of tasks Two parts – oral and written – with their own administration guidelines	More effective teaching in the classroom (a longer term outcome)
Administration	Vague administration procedures – from timing to marking (◀ Chapter 5; Chapter 7; Chapter 8)	Clear set of assessment criteria Structured timing Examiner behaviour specified Clear marking guidelines	Clearer understanding of expected performance for students; clearer guidelines for preparation Easier marking for examiners; marking system becomes more transparent for students More transparent feedback for students
Validity	Content of the Exam disagrees with modern views on teacher competence Exam tasks have no relation to teacher job	Assessment focus is brought closer to expected teacher performance (◀ Chapter 3, 4; Chapter 10, 11)	More valid Final Language Examination which is compatible with international experience
Reliability	Serious threats to reliability: unequal conditions for exam takers Possibility of cheating Vague assessment criteria (◀ Chapter 7; Chapter 8)	Fewer threats to reliability: clearer assessment criteria clearer administration guidelines (timing, examiner behaviour, marking)	More reliable Final Language Examination, compatible with international examinations and agreeing with innovations in the Russian education system (◀ Chapter 2)
Exam design	An internal decision of the department Tasks designed at the Department No moderation or piloting No external opinion on designed tasks Choices of tasks are often convenience choices	School teachers and/or other stakeholders [possibly] involved in moderation and piloting Clearer guidelines on task and text selection Clearer guidelines on expected performance	More stakeholders involved More transparency achieved More efficient procedure for task design

Changes in the Final Language Examination would be beneficial, first of all, to the Exam itself by enhancing its validity, reliability, authenticity and practicality. Renewed content and format, together with revised administration and marking procedures would bring the Examination closer to positive experiences of other countries (e.g. USA, Australia, Hong Kong) and international examinations for teachers (e.g. Cambridge ESOL). Having assessment foci in good agreement with national and international standards is beneficial in terms of coherence and, therefore, the availability of various educational systems to university graduates. However, at the moment the suggested changes can be treated only as recommendations because they require, first of all, approval of various stakeholders and substantial effort on their part. Any changes, if decisions are taken to introduce them, would require time and involve several teams of professionals.

13.3. Limitations of this study

One of the major limitations of this research is its quite narrow scale, with the case study involving only one institution (Tula State Pedagogical University) in one region of Russia (Tula region). However, there are some reasons to suggest that the situation observed in Tula is quite typical of that in other regions and teacher training institutions in Russia, and even in some former USSR republics (e.g. Belarus). The British Council project Tula PRESET (1998-2007), referred to in Chapter 2, aimed to collect data from various regions of Russia on the content and format of continuous and final assessment at university. The data provided by 5 regions⁷¹ demonstrated that assessment formats were very similar to the one under current study – a linguistic question and mostly receptive skills, although speaking was also a part of the final examinations. No further data collection was performed, so a wider study of final language assessment in Russia could become a direction for further research, or for an independent research. This research is seen as an initial step in the investigation of language assessment for language teachers in Russia, and the scale of data collection was deliberately kept quite narrow (covering one region only). A detailed picture was required for suggesting any

⁷¹ The regions participating in the Project were Ekaterinburg, Krasnoyarsk, Nizhny Novgorod, Omsk, Volgograd

possible changes in the Exam under study and this seemed possible in a limited context.

Evaluation of the Exam is based on opinions of different stakeholders – from Exam designers to school teachers of English who either had passed such an exam themselves, or deal with younger teachers who did, or both. Such data collection resulted in quite a detailed picture and multi-faceted analysis of the current situation. At the same time, no external evaluation was employed, such as, for example, institutional exam validation conducted by an independent evaluation body. This could make analysis of Exam strengths and weaknesses more reliable. Nevertheless, there is currently no such body in Russia that would evaluate assessment instruments, and such a procedure is not expected to take place (see Chapter 2).

This research studies current issues of Exam design without looking at the Exam from a historical perspective. In recent years the current Exam has undergone some changes, but the changes were mostly cosmetic, and it can be claimed that the format under study has been employed for at least 20 years. This research could have benefited from interviews with senior faculty teaching staff members⁷², a survey, or a discussion group, which could have cast some light on why some decisions had been made and why the current format has been considered optimal for future teachers (university graduates). Such data could have been useful in discussing possible changes in the current Examination because it might have explained some choices previously made, the reasons for those choices and, possibly, some theories underlying them.

Design of alternative Exam tasks was based on the analysis of data from the stakeholders and documents, with every effort made to take into consideration all facts and opinions. However, no piloting or moderation of alternative tasks was performed, for various reasons, the major ones being time constraints on this research and the need to involve a considerable number of people in Russia to serve as a piloting population. Piloting could have produced more evidence as to how the current Examination is seen by various groups of stakeholders. Therefore, piloting of the alternative Exam tasks is seen as a further stage of research (see Conclusion).

⁷² There are staff members at the Faculty of FL with work experience of 40 years and more

Conclusion

This study investigates methods of final language assessment for teachers of English as a Foreign Language (university graduates) in the Russian educational context. Final Language Assessment at university takes place in the form of the Final Language Examination at the end of the course of studies (4th/5th year). The Examination can be considered a high-stakes examination because passing it is an essential step for graduates in obtaining their teacher qualification. The research aims to evaluate the current form of the Final Language Examination for university graduates who are future teachers of English in Russia, and to suggest possible changes in the Exam's content, format and administration.

Language development has been seen as a key element of foreign language (FL) teacher training programmes in various countries, especially where the job of FL teaching is done by a non-native speaker (e.g. Medgyes, 1999; Ćurković-Kalebić, 2004; Sešek, 2007; Coniam, 2013). Within an abundant range of publications on FL teaching not much, however, has been written about language courses for language teachers (e.g. Spratt, 1996; Thornbury, 1997). As Sešek (2007) observed, quite often teacher's language development programmes focus on general language and English literature, without paying enough attention to professional (teachers') language courses.

Similarly to general language teaching that aims to develop language competence (e.g. Common European Framework, 2001), language development of (future) FL teachers brings in the issue of language teacher language competence, its structure and key elements. Although the concept of 'language teacher language competence' seems to be a key concept in language teacher education, not much has been written that provides a clear definition of it (e.g. Kennedy, 1983; Thomas, 1987; Widdowson, 2002; Richards, 2010). This research sees language teacher language competence development as a combination of various interrelated elements:

- linguistic competence
- teacher language awareness
- communicative skills of listening, reading, speaking, writing

- Classroom Language, and also pedagogical knowledge (Richards, 2010) and knowledge and practical command of TESOL.

Assessment of language teacher language competence follows the principles of general language assessment, as seen by Alderson (1995), Hughes (2003), H.D.Brown (2004), McNamara (2006), and Bachman&Palmer (2010):

- validity (content, construct, concurrent, predictive)
- reliability of administration and marking procedures
- authenticity of assessment input (tasks) and output (expected performance)
- practicality.

However, as pointed out by Grant (1997), Elder (2000) and Consolo (2008), assessment of teacher language competence has several features that make it different from general language assessment. One difference can be seen quite distinctly – it is the content of assessment tasks that might differ for general and teacher assessment. Apart from this, tests for language teachers are often expected to be direct, as opposed to the majority of indirect general language tests, i.e. to be administered at the workplace, but not in the examination room (McNamara, 1997). On the one hand, the direct mode of assessment enhances its authenticity, on the other hand, according to Grant (1997) and Elder (2000), it adds some threats to the validity and reliability of tests administered directly. Therefore, one of the key issues in language teacher assessment is finding an optimal balance between validity, reliability, authenticity and practicality by means of an adequate choice of tasks, content and mode of an administration. Examples of these can be found in some national (e.g. LPATE) and international (e.g. TKT, ICELT, Praxis®) language examinations for language teachers that were developed within the last 15 years.

The Final Language Examination for university graduates (future teachers of English) in Russia can be considered a version of a language test for language teachers. The high-stakes status of this Examination makes its quality in general and its validity and reliability in particular crucial for both test takers and the university. In the situation under study, the Examination is designed internally, with no piloting or external evaluation, which is considered one of the essential steps in test design (e.g. Alderson, 1995; Hughes, 2003; H.D.Brown, 2004).

To evaluate the current Final Language Examination, research instruments – an interview framework and 3 questionnaires - were developed and used with different groups of stakeholders: Exam developers from the university, Exam takers, and school teachers of English. The collection of empirical data that was conducted in 2012-2013 involved about 150 participants in interviews and surveys administered face-to-face and online.

The obtained data revealed some positive features of the Examination under study: well-established materials design procedures, a team of experienced examiners, a correlation between the Exam content and the language content of the programme of studies. However, along with these advantages, some problematic areas were disclosed in the current Final Language Examination. The areas that cause most concern are:

- *validity* of the Exam tasks and lack of any direct relationship between what the Exam assesses and what teachers of English require in and out of the language classroom;
- *reliability* of the Exam administration, especially unspecified examiner behaviour and threats to rater reliability caused by a vague system of assessment criteria and unclear definition of the expected performance of Exam takers;
- *authenticity* of the Exam tasks that do not represent the situations that teachers of English face in their workplace.

Apart from evaluation of the current Final Language Examination for university graduates in Russia, this research also suggests possible changes in the Exam content, format and administration by presenting alternative Exam tasks and assessment criteria. It is hoped that such changes may contribute to Exam validity, reliability and practicality. The suggested changes can be presented in 3 groups.

Changes in the *Examination content* involve re-considering the assessment foci, designing a detailed taxonomy of skills under assessment, widening the range of topics by including both general and professional English, and widening the range of input texts for receptive and productive tasks. The 2nd group of changes – changes in the *Examination format* - might start with widening the range of Exam task types. More global changes might include adding new parts to the Exam (e.g. Writing) or significant re-consideration of the existing parts (e.g. Task 1: Linguistic knowledge).

Changes in the Exam content and format would make unavoidable the 3rd group of changes – in the *Exam administration*. This could, first of all, include design of a detailed assessment scale and descriptors, and also a thorough description of assessment performance.

Another set of changes may be seen as a part of Exam design or a dimension of Exam administration. Training for Exam designers and raters is considered crucial in contemporary research (e.g. Alderson, 1995; Lumley, 2002; Brown, 2004; Bachman & Palmer, 2010; Kuiken&Vedder, 2014; Yan, 2014). Adequate training for examiners is seen as a means of improving rater reliability, whereas training for Exam designers might contribute both to validity and reliability of Exam tasks.

The major outcomes of this research are seen, first of all, in the thorough evaluation of the existing Final Language Examination for university graduates (future teachers of English) in Russia. Although language development of future teachers has always been considered one of the priorities of FL teacher training (e.g. Schukina, 2008), no research has been conducted in the area of teacher pre-service or in-service language assessment. Besides analysing strengths and weaknesses of the current Final Language Examination, the evaluation done in this research allows for suggesting alternative methods of the Final Language Exam's design and administration. Such suggestions can be considered beneficial in the situation under study and in the wider context of other Russian teacher training institutions.

Another outcome of this research is design of Exam evaluation instruments:

- a questionnaire for Exam designers that focuses on Exam design procedures, selection of Exam content and tasks, and Exam administration and marking procedures;
- post-exam questionnaires for examiners and Exam takers, focusing on Exam administration;
- a Needs Analysis questionnaire for language teachers at school/college/university and an online Needs Analysis questionnaire for final year university students, in which the respondents are asked to indicate what language skills they employ in and out of the classroom;
- a structured interview framework for practising language teachers that involves respondents in speculation on Exam task appropriacy and relevance to the teacher job.

The outcomes of this research suggest possible ways of developing in several directions. First, it may develop into a full-scale research on language teacher language competence, involving more participants from more regions and, probably, several countries. Such research may result in a detailed description of teacher language competence and its elements, a list of professional vocabulary items, and a more precise description of teacher language awareness, which would represent the development of ideas expressed by Widdowson (2002), Wright (2002), and Bolitho&Carter (2003). A description of language competence might also include description of its levels and expected performance in a way similar to the description of general language competence presented in the Common European Framework of Reference (2001). Such research might be in good agreement with the European Portfolio for Student Teachers of Languages (2007), which concentrates mostly on teacher competence in general and only gives an outline of the language skills a FL teacher is expected to demonstrate.

Another direction might be a search for deeper insight into final language assessment at universities in different parts of Russia. This might involve up to 100 universities that train FL teachers and could come up with an analysis of a variety of assessment formats, approaches to exam administration, and choices of content. On a larger scale, such research might involve universities from other countries. Comparison of final assessment for language teachers in various countries (e.g. former USSR, or former socialist countries – Poland, Bulgaria, Romania, Hungary, and countries of the former Yugoslavia) might be interesting not only from the point of view of implementation of assessment principles, but also as implementation of educational policies in countries with once similar ideological and political backgrounds.

As a variation of the direction discussed above, which aims to investigate assessment systems in Russian pedagogical universities, research might be undertaken to further develop alternative assessment materials and their piloting, followed by their implementation in one university or several. Such work might lead to changes in the whole system of assessment at the pedagogical university, including continuous assessment.

List of references

- Akyel, A.S., Ozek, Y. (2002). A language needs analysis research at an English medium university. *Procedia Social and Behavioral Science*, 2, 969-975.
- Alderson, C. (2000). *Assessing reading*. Cambridge University Press.
- Alderson, J. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- American Council on the Teaching of FL. (2002). *Program Standards for the preparation of FL teachers*.
- Anastasi, A. (1988). *Psychological testing*. New York: Macmillan Publishing Company.
- Andrews, S. (2002). Teacher language awareness and language standards. *Journal of Asian Pacific Communication*, 12(1), 39-62.
- Australian Association of Modern Languages Teachers. (2005). *Final Report on the Development of Standards for Teachers of Indonesian Project*.
- Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.
- Bachman, L., Palmer, A. (2010). *Language Assessment in Practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Bachman, L.F., Palmer, A.S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.
- Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL iBT writing tasks. *Language Testing*, 31(2), 241-259.
- Berry, R. (1990). The Role of Language Improvement in In-service Teacher Training: Killing Two Birds with One Stone. *System*, 18(1), 97-105.
- Bolitho, R., Carter, R., Hughes, R. (2003). Ten Questions about Language Awareness. *ELT Journal*, 57(3), 251-259.
- Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. New York: Longman.
- Brown, J. (2000). What is construct validity? *JALT Testing and Evaluation Newsletter* 4(2), 8-12.
- Chalhoub-Deville, M. (1997). Theoretical Models, Assessment Frameworks and Test Construction. *Language Testing*, 14(1), 3-22.
- Chapelle, C. (2011). Validity argument for language assessment: the framework is simple... . *Language Testing*, 29(1), 19-27.
- Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English*, 65-81.
- Cohen, L., Manion, L., Morrison, K. (2007). *Research Methods in Education*. Routledge.

- Coniam, D., Falvey, P. (2002). Selecting models and setting standards for teachers of English in Hong Kong. *Journal of Asian Pacific Communication*, 12(1), 13-37.
- Coniam, D., Falvey, P. (2013). Ten years on: the Hong Kong Language Proficiency Assessment for Teachers of English (LPATE). *Language Testing*, 30(1), 147-155.
- Consolo, D. A., Porto, C. F. C. (2013). Teachers' Competences from Foreign Language Teachers' Perspectives. *Revista SOLETRAS*, 26, 1-15.
- Consolo, D.A., Alvarenga, M.B., Concario, M. (2009). An examination of foreign language proficiency for teachers (EPPL): The initial proposal and implications for the Brazilian context. *The Teaching of English: Towards an Interdisciplinary Approach Between Language and Literature* (pp. 1-15). São José do Rio Preto-SP: ABRAPUI.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Cambridge University Press. Retrieved from www.coe.int/lang
- Coyle, D., Hood, P., Marsh, D. (2010). *CLIL: content and language integrated learning*. Cambridge: Cambridge University Press.
- Creswell, J. (2003). *Research Design. Qualitative, quantitative and mixed methods approaches*. SAGE Publications.
- Cullen, R. (1994). Incorporating a Language Improvement Component in Teacher Training Programmes. *ELT Journal*, 48(2), 162-172.
- Čurković-Kalebić, S. (2005). Prema razvoju standarda u obrazovanju budućih nastavnika stranoga jezika. *ATTE 30th Annual Conference*. Amsterdam.
- Dalton-Puffer, C., Nikula, T. (2014). Guest editorial: content and language integrated learning. *The Language Learning Journal*, 42(2), 117-122.
- Davies, A. (2010). Test fairness: A response. *Language Testing*, 27(2), 171-176.
- Davies, A. (2012). Kane, validity and soundness. *Language Testing*, 29(1), 37-42. Retrieved January 7, 2015
- Department for Education. (2011). *Teachers' Standards. Guidance for school leaders, school staff and governing bodies*. Retrieved July 5, 2014
- Douglas, D. (2002). *Assessing Languages for Specific Purposes*. New York: Cambridge University Press.
- Douglas, D. (2014). *Understanding Language Testing*. Routledge.
- Edge, J. (1988). Applying Linguistics in English Language Teacher Training for Speakers of Other Languages. *ELT Journal*, 42(1), 9-13.
- Elder, C. (1994). Performance testing as a benchmark for LOTE teacher education. *Melbourne Papers in Language Testing*, 3(1), 1-25.
- Elder, C. (2001). Assessing the Language Proficiency of Teachers: Are There Any Border Controls? *Language Testing*, 18(2), 149-170.

- Ellis, V. (2007). *Subject Knowledge and Teacher Education. The development of Beginning Teachers' Thinking*. Continuum.
- ETS. (2012). *The Praxis Study Companion*. Princeton. Retrieved November 5, 2014, from <https://www.ets.org/s/praxis/pdf/5195.pdf>
- Ferris, D. (2012). Writing Instruction. In J. Richards, *The Cambridge guide to pedagogy and practice in second language teaching* (pp. 226-237). Cambridge University Press.
- Flick, U. (2009). *An Introduction to Qualitative Research*. London: Sage.
- Fowler, F. (1993). *Survey Research Methods*. SAGE Publications.
- Fulcher, G., Davidson, F. (2006). *Language Testing and Assessment: An advanced resource book*. Routledge.
- Gablasova, D. (2014). Issues in the assessment of bilingually educated students: expressing subject knowledge through L1 and L2. *The Language Learning Journal*, 42(2), 151-164.
- Georgiou, S. (2012). Reviewing the puzzle of CLIL. *ELT Journal*, 66(4), 495-504.
- Grant, L. (1997). Testing the Language Proficiency of Bilingual Teachers: Arizona's Spanish Proficiency Test. *Language Testing* 14 (1), 23-46.
- Hamilton, L. (2013). *Using case study in education research*. London: SAGE.
- Hamp-Lyons, L. (1996). The challenges of second-language writing assessment. In E. L. White, *Assessment of writing: politics, policies, practices* (pp. 226-240). New York: Modern Language Association of America.
- Hamp-Lyons, L. (1997). Washback, Impact and Validity: Ethical Concerns. *Language Testing*, 14(3), 295-303.
- Harmer, J. (2007). *The Practice of English Language Teaching (4th Edition) (With DVD)*. Longman.
- Heaton, J. (1995). *Writing English Language Tests*. New York: Longman.
- Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Hughes, G. (1987). *A handbook of Classroom English*. Oxford University Press.
- Huhta, A., Alanen, R., Tarnanen, M. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing*, 31(3), 307-328.
- Huisman, J., van der Wende, M. (2004). The EU and Bologna: are supra- and international initiatives threatening domestic agendas? *European Journal of Education*, 39(3), 349-357.
- Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. H. Williamson, *Validating holistic scoring for writing assessment: theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

- Hutchinson, T., Waters, A. (1990). *English for Specific Purposes. A learning-centred approach*. Cambridge University Press.
- Irkutskaya, V. (2011). WTO and modernization of Russian higher education system. *Вестник ТГПУ*.
- Jasso-Aguilar, R. (2015). Sources, Methods and Triangulation in Needs analysis: A critical perspective in a case study of Waikiki hotel maids. In M. Long, *Second Language Needs Analysis* (pp. 127-158). Cambridge University Press.
- Kane, M. (2010). Validity and Fairness. *Language Testing*, 27(2), 177-182.
- Kane, M. (2012). Validating score interpretations and uses: Messick lecture, Language Testing Research Colloquium, April 2010. *Language Testing*, 29(1), 3-17.
- Kelly, M. G. (2004). *European Profile for Language Teacher Education: a frame of reference*. Council of Europe, Strasbourg.
- Kennedy, C. (1983). An ESP Approach to EFL/ESL Teacher Training. *The ESP Journal*, 2, 73-85.
- Kennedy, C., Bolitho, R. (1984). *English for Specific Purposes*. Macmillan.
- Knoch, U. (2010). Investigating the effectiveness of individualized feedback to rating behaviour – a longitudinal study . *Language Testing* , 28(2), 179-200.
- Kuiken, F., Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, 31(3), 279-284.
- Kuiken, F., Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, 31(3), 329-348.
- Language Education Policy Profile for Poland. (2005). Council of Europe; Ministerstwo Edukacji Narodowej.
- Lavigne, A. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers and students. *Teachers College Record*, 116(1). Retrieved March 7, 2016, from <http://www.tcrecord.org>
- Ling, G., Mollaun, P., Xi, X. (2014). A study of the impact of fatigue on human raters when scoring speaking responses. *Language Testing*, 31(4), 479-499.
- Lowe, M. (2007). *Beginning Research*. Routledge.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3), 246-276.
- Lynch, B. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.
- Malchenko, A., Lyubimova, Z., Okninskaya T. (1995). *English for Primary School Teachers*. Tula: Interbumaga.
- McNamara, T. (2000). *Language Testing*. Oxford University Press.
- Medgyes, P. (1999). *The Non-native Teacher*. Hueber.
- Messick, S. (1989). Validity. In R. Linn, *Educational Measurement* (pp. 13-103). New York: Macmillan.

- Neave, G. (2003). The Bologna Declaration: Some of the Historic Dilemmas Posed by the Reconstruction of the Community in Europe's Systems of Higher Education. *Educational Policy*, 17(1), 141-164.
- Newby, D., Allan, R., Fenner, A-B. (2007). *European Portfolio for Student Teachers of Languages: A reflection tool for language teacher education*. Graz: Council of Europe.
- Norris, J. (2008). *Validity Evaluation in Language Assessment*. Frankfurt am Mein: Peter Lang GmbH.
- Nunan, D. (1995). *Research Methods in Language Learning*. Cambridge: Cambridge University Press.
- Nunan, D., Carter, R. (Eds). (2001). *The Cambridge Guide to Teaching English to Speakers of Other Languages*. Cambridge: Cambridge University Press. doi:<http://dx.doi.org/10.1017/CBO9780511667206>
- Parrott, M. (1993). *Tasks for Language Teachers. A Resource Book for Training and Development*. Cambridge: Cambridge University Press.
- Parrott, M. (2010). *Grammar for English Language Teachers*. Cambridge: Cambridge University Press.
- National Council for Accreditation of Teacher Education (2002). *Program Standards for the Preparation of FL Teachers*. Retrieved 2010
- Richards, J. (2008). Second language teacher education today. *RELC Journal*, 39(2), 158-176. Retrieved February 24, 2011, from www.professorjackrichards.com
- Richards, J. (2010). Competence and Performance in Language Teaching. *RELC Journal*, 41(2), 101-122. Retrieved February 20, 2011, from www.professorjackrichards.com
- Richards, J.C., Platt, J., Platt, H. (1997). *Dictionary of Language Teaching and Applied Linguistics*. Harlow: Longman.
- Richards, J.C., Renandya, W.A. (Eds). (2002). *Methodology in Language Teaching. An Anthology of Current Practice*. Cambridge: Cambridge University Press. doi: <http://dx.doi.org/10.1017/CBO9780511667190>
- Scrivener, J. (2010). *Learning Teaching*. Macmillan.
- Sešek, U. (2007). English for Teachers of EFL - Toward a holistic description. *English for Specific Purposes*, 26(4), 411-425.
- Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11(2), 99-123.
- Skuja, R. M. (1994). English for teaching purposes - a Singapore experience. In V. Barkley, *Future directions in English Language Teacher Education. Asia and Pacific perspectives* (pp. 161-173). Hong Kong: Institute of Language in Education.
- Spratt, M. (1994). *English for the Teacher. A Language Development Course*. Cambridge: Cambridge University Press.

- Spratt, M., Pulverness, A., Williams, M. (2006). *The TKT Course*. Cambridge: Cambridge University press.
- Thomas, A. (1987). Language Teacher Competence and Language Teacher Education. In R. Bowers, *Language Teacher Education: An Integrated Programme for EFL Teacher Training* (pp. 33-42). Modern Language Publications.
- Thornbury, S. (1997). *About Language - Tasks for Teachers of English*. Cambridge: Cambridge University Press.
- Thornbury, S. (2006). *An A-Z of ELT*. Macmillan.
- Tomlinson, P., Saunders, S. (1995). The Current Possibilities for Competence profiling in Teacher Education. In A. Edwards, *Assessing Competence in Higher Education* (pp. 79-97).
- Trappes-Lomax, H. (2002). Language in Language Teacher Education: a discourse perspective. In H. Trappes-Lomax, *Language in Language Teacher Education* (pp. 1-19). Amsterdam: Philadelphia.
- University of Cambridge, ESOL Examinations. (2005). *ICELT. In-service Certificate in English Language Teaching. Syllabus and Assessment Guidelines*. Cambridge. Retrieved July 17, 2014, from <http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/icelt/>
- University of Cambridge, Language Assessment. (2010). *TKT all modules. Teacher Knowledge Test*. Cambridge. Retrieved July 15, 2014, from www.cambridgeenglish.org/tkt
- University of Cambridge, Language Assessment. (2011). *CELTA. Certificate of English Teaching to Speakers of Other Languages*. Cambridge. Retrieved July 20, 2014, from <http://www.cambridgeenglish.org/images/celta-brochure-2013.pdf>
- University of Cambridge, Language Assessment. (2011). *DELTA. Diploma in Teaching English to Speakers of Other Languages*. Cambridge. Retrieved July 14, 2014, from <http://www.cambridgeenglish.org/teaching-english/teaching-qualifications/delta/>
- University of Cambridge, Language Assessment. (2014). *Summary of Changes to Delta Module One Examination for 2015*. Cambridge. Retrieved November 5, 2014, from <http://www.cambridgeenglish.org/images/174149-summary-of-changes-to-delta-module-one-examination-for-2015.pdf>
- Ur, P. (1999). *A Course in Language Teaching*. Cambridge: Cambridge University Press. Retrieved from <http://ebooks.cambridge.org/ebook.jsf?bid=CBO9780511732928>
- van Ek, J., Trim, J.L. (1998). *Threshold level*. Cambridge University Press.
- van Ek, J., Trim, J.L. (1998). *Waystage level*. Cambridge University Press.
- Weigle, S. (2002). *Assessing writing*. Cambridge University Press.
- Widdowson, H. (2002). Language Teaching: defining the subject. In H. Trappes-Lomax, *Language in Language Teacher Education* (pp. 67-81). Amsterdam: Philadelphia.

- Wideen, M.F., Grimmett, P.P. (1995). *Changing Times in Teacher Education. Restructuring or Reconceptualisation?*
- Wiersma, W., Jurs, S.G. (2005). *Research Methods in Education*. Pearson Education.
- Willis, J. (1987). *Teaching English through English*. Longman.
- Wright, T. (2002). Doing Language Awareness: issues for language study in language teacher education. In H. Trappes-Lomax, *Language in Language Teacher Education* (pp. 113-130). Amsterdam: Philadelphia.
- Wright, T. (2010). Second Language Teacher Education: Review of recent research on practice. *Language Teaching*, 43(3), 259-296.
- Wright, T., Bolitho, R. (1993). Language Awareness: A Missing Link in Teacher Education. *ELT Journal*, 47(4).
- Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, 27(2), 147-170.
- Акулич, М. (2005). Образование в условиях глобализации. *Университетское управление: практика и анализ*.
- Гретченко, А. (2006). Автономизация высшего образования и Болонский процесс. *Высшее образование*, 6.
- Гретченко, А. (2009). Двухуровневая система высшего образования: европейский опыт. В монографии «Образование. Экономика. Финансы. Модернизация». М: ВГНА Минфина России.
- Гретченко, А. (2011). *Перспективы развития высшего профессионального образования в России*. М: МГППУ.
- Иркутская, В. (2011). ВТО и модернизация системы высшего образования в России. *Вестник Томского государственного педагогического университета*, 6.
- Кузьмина, Л.Г., Стернина М.А. (2009). О качестве интернет-тестирования ФЭПО по английскому языку. *Вестник ВГУ. Лингвистика и межкультурная коммуникация*, 1.
- Мальченко А.А., Соколова Н.Г. (2002). *Дневник работы студента-практиканта отделения заочного обучения факультета иностранных языков: Учебно-методическое пособие для студентов ОЗО педагогических вузов*. Тула: Изд-во Тул. гос. пед. ун-та им. Л.Н. Толстого.