

UNIVERSITY OF BELGRADE

FACULTY OF PHILOLOGY

Milan Milanović

INVESTIGATING AUTHENTIC FORMS OF
ASSESSMENT IN TESTING ENGLISH FOR
SPECIFIC PURPOSE SPEAKING SKILLS

Doctoral Dissertation

Belgrade, 2019

УНИВЕРЗИТЕТ У БЕОГРАДУ

ФИЛОЛОШКИ ФАКУЛТЕТ

Милан Милановић

ИСПИТИВАЊЕ АУТЕНТИЧНИХ ОБЛИКА
ПРОВЕРЕ ЗНАЊА У ТЕСТИРАЊУ
ГОВОРНИХ ВЕШТИНА НА ЕНГЛЕСКОМ
ЈЕЗИКУ СТРУКЕ

докторска дисертација

Београд, 2019.

БЕЛГРАДСКИЙ УНИВЕРСИТЕТ
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

Милан Милановић

ИЗУЧЕНИЕ АУТЕНТИЧНЫХ ФОРМ
ОЦЕНИВАНИЯ НАВЫКОВ ГОВОРЕНИЯ ПРИ
ТЕСТИРОВАНИИ АНГЛИЙСКОГО ДЛЯ
ОСОБЫХ ЦЕЛЕЙ

докторская диссертация

Белград, 2019.

МЕНТОР:

проф. др Оливера Дурбаба, редовни професор

Филолошки факултет

Универзитет у Београду

ЧЛАНОВИ КОМИСИЈЕ ЗА ОЦЕНУ И ОДБРАНУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ:

ДАТУМ ОДБРАНЕ РАДА: _____

Investigating authentic forms of assessment in testing English for specific purpose speaking skills

Abstract

This doctoral dissertation has attempted to investigate authentic forms of assessment in testing ESP speaking skills. To achieve this objective, specific purpose target language use speaking tasks were identified in collaboration with subject specialist informants and by the means of context-based qualitative research, helping the researcher extract speaking task characteristics in the real life domain. The identified domain is that of a labor market in which Business English is used as a language of communication in companies registered at the territory of Kragujevac (Sumadija and Pomoravlje County, Serbia). The researcher analyzed English language speaking tasks by the means of Task characteristics framework, which enabled him to emulate the characteristics of the speaking tasks, embedding them into the characteristics of speaking test tasks. By utilizing the Task characteristics framework, the researcher developed speaking test tasks which claim enhanced situational and interactional authenticity compared to less contextualized speaking tasks, developed by following a syllabus-based model of construct definition. These newly developed tasks were presented in a series of formative assessments to a group of 150 business students, enrolled in three different modules at the Faculty of Economics (University of Kragujevac), along with other aspects of authentic assessment – self-evaluation, peer-evaluation, and feedback. The results obtained by assessing students' performance were collected and subjected to statistical analyses for the purpose of finding answers to the following research questions: (1) Can target language use situation tasks be used as a model for authentic classroom test tasks? (2) Do authentic forms of assessment exert a positive influence on students' progress? (3) Should background knowledge be tested in specific purpose speaking assessments? (4) Do authentic forms of assessment exert a positive influence on students' awareness of their own progress? (5) Do business students possess the language skills matching the needs of the labor market? To find answers to these research questions the author formulated and tested the following hypotheses: (1) The examinees who have been thoroughly trained to apply evaluation criteria demonstrate a better overall performance in the final oral exam in comparison to the examinees who have not been thoroughly trained on applying analytic and holistic scoring criteria in assessing their own and the performance of their peers; (2) Performing

on a task requiring that test takers should possess background knowledge related to the field of *Marketing*, the Control group demonstrates very similar results to the more successful of the two experimental groups; (3) End of semester survey results indicate that more than two thirds of the examinees demonstrate positive perceptions of authentic tasks, as well as of the system of evaluation and self-evaluation that they have been exposed to; (4) End of semester self-evaluation questionnaire results indicate that at least 70% of the Control group's responses provided to estimate their target skills match the responses provided at the beginning of the semester; (5) End-of-semester self-evaluation results indicate that at least half of the sample in the Experimental groups achieved progress by one CEFR level, as corroborated by the Second placement test results; and (6) The highest agreement in responses to the "Can-do" survey is the one between subject specialist informants and Group 1 subjects. The analysis of the research results helped the author find answers to research questions and reach the following conclusions: (1) TLU speaking tasks can be used as a model for designing authentic tasks for classroom use, following a thorough analysis of the context in which target language use occurs. Two methods are recommended to this end: context-based research and grounded ethnography, performed in collaboration with subject specialist informants. The resulting set of task characteristics is used as a model for test task characteristics, sharing situational and interactional authenticity with test tasks. (2) In response to the second research question, the author conducted an empirical research with subjects exposed to authentic test tasks, within the task-based approach to assessment, by which task deliverables had relevance to the TLU contexts. In addition, the subjects were familiar with evaluation criteria and took accountability for the learning outcomes that the assessment was linked to. The results confirm that students' exposure to authentic test tasks and methods of evaluation and self-evaluation has a positive impact on students' progress, as corroborated by their achievement in summative assessments. (3) In response to the third research question, the research results suggest that background knowledge exerts a positive influence on task achievement, even with weaker learners, helping them alleviate accuracy-related deficiencies while attending to the task. (4) One of the objectives of this study was to investigate if authentic forms of assessment exert a positive influence on students' awareness of their own progress. Research results indicate that students' perceptions of assessment methods play vital role in their engagement in the task, and consequently in their progress. In addition, students who are trained in monitoring and rating their own progress demonstrate a better overall

success in both formative and summative assessments. (5) The needs analysis conducted prior to the commencement of the research indicated that there was a discrepancy between the English language skills that university degree holders possessed and the actual language needs in the labor market. The empirical part of the research proved that when students are continuously exposed to authentic language tasks, as well as to authentic forms of assessment and self-assessment, their language performance stands in line with labor market requirements. The study presented in this doctoral dissertation makes several contributions to theory and practice of language assessment. First, it contributes to a better understanding of speaking assessment. Second, it promotes a process of test development that takes into consideration situational and interactional authenticity of speaking tasks. Third, it offers methodology for ensuring that discrepancy between the realms of academia and the real world is minimized. Fourth, the study makes methodological contributions to test task analysis and development. Fifth, the study has pedagogical relevance in that it advocates student-centered learning and testing. Finally, it results in a number of recommendations relevant to curricular amendments at the Faculty of Economics, University of Kragujevac.

Key words: assessment, authenticity, ESP, task, Business English, construct, target language use, assessing speaking.

Scientific field: Linguistics

Scientific subfield: applied linguistics, testing

UDK number: _____

Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику струке

Апстракт

Аутор ове докторске дисертације узео је за предмет истраживања тестирање говорних вештина на енглеском као језику струке, настојећи да истражи аутентичне облике задатака којима се овај језик тестира. Како би остварио постављене циљеве, аутор је сарађивао са стручњацима из посматране области да би идентификовао говорне задатке који се односе на употребу енглеског, као језика струке. Посредством контекстуализованог квалитативног истраживања, аутор докторске дисертације екстраховао је задатке са којима се говорници суочавају у тзв. „домену стварног живота“ и пренео их у образовни домен, сачувавши њихове најважније карактеристике. За домен „стварног живота“ узет је домен тржишта рада, који је додатно сужен на компаније у којима се пословни енглески језик користи као језик пословне комуникације, на територији Крагујевца (у оквиру шумадијско-поморавског региона у Републици Србији). Говорни задаци који се решавају посредством енглеског језика анализирани су употребом Оквира карактеристика задатака, захваљујући коме се карактеристике говорних задатака ван образовног домена преносе у тај домен са минималним осипањем основних обележја задатака из домена стварног живота. На тај начин, новонастали тестовни задаци претпостављају виши степен ситуационе и интеракцијске аутентичности него што је то случај код тестовних задатака са слабијом контекстуализацијом, односно оних који су изведени на основу дефиниције конструкта настале на основу силабуса. Тестовни задаци којима се проверава вештина говора на енглеском језику састављени су на основу горе поменутог Оквира и представљени групи од 150 студената економије, који су уписани на три различита модула на Економском факултету Универзитета у Крагујевцу. Осим тестовних задатака, испитаницима су представљени и други аспекти аутентичних облика тестирања, попут самоевалуације, евалуације вршњака, и давања/добијања повратне информације. Резултати настали евалуацијом постигнућа студената који су учествовали у студији повргнути су статистичким анализама са циљем проналажења одговора на следећа истраживачка питања: (1) Да ли задаци настали у ситуацијама у којима се употребљава циљни језик могу да послуже као модел за тестовне задатке у образовању? (2) Да ли

аутентични облици испитивања имају позитиван утицај на постигнуће студената? (3) Да ли предзнање треба да буде предмет тестирања у испитивању вештине говора у случају енглеског језика за посебну намену? (4) Да ли аутентични облици тестирања врше позитиван утицај на свест студената о сопственом напретку? (5) Да ли студенти економије поседују језичке вештине које одговарају потребама тржишта рада? Аутор рада поставио је следеће хипотезе како би пронашао одговоре на горе поменута питања: (1) Испитаници који су детаљно обучавани да примењују критеријуме за евалуацију постижу бољи успех на завршном усменом испиту у односу на испитанике који нису прошли детаљну обуку за примену аналитичке и холистичке рубрике приликом оцењивања сопственог и постигнућа вршњака; (2) Приликом извршења задатка који подразумева предзнање из области маркетинга, испитаници из контролне групе остварују приближно исте резултате као испитаници из успешније експерименталне групе; (3) Резултати анкете спроведене на крају семестра указују на то да више од две трећине испитаника има позитивне ставове према аутентичним задацима, као и облицима само-евалуације и евалуације вршњака; (4) Анализа резултата упитника који се односи на идентификацију циљних језичких вештина указује на то да се најмање 70% одговора које су испитаници контролне групе дали на крају семестра поклапа са одговорима датим на почетку семестра; (5) Резултати самоевалуације спроведене на крају семестра указују на то да је најмање половина узорка у експерименталним групама достигла напредак за један језички ниво ЗЕРОЈ-а, што је потврђено и резултатима другог класификационог теста; и (6) Највеће подударње у одговорима датим приликом спровођења „Can-do“ анкете постоји између стручњака из привреде и испитаника из Групе 1. Након анализе добијених резултата, аутор је дошао до следећих закључака: (1) задаци настали у ситуацијама у којима се употребљава циљни језик могу да послуже као модел за аутентичне тестовне задатке уколико се израђују на основу детаљне анализе контекста у коме настају ван образовног домена. Студија издваја две корисне методе уз помоћ којих се анализа задатака врши са успехом: анализа контекста и метода „утемељене етнографије“; обе у сарадњи са стручњацима из одговарајућих области. Захваљујући овим методама, састављачи тестова добијају скуп карактеристика језичких задатака који су ситуационо и интеракцијски аутентични са тестовним задацима који се касније употребљавају у контексту образовања. (2) У потрази за одговором на друго истраживачко питање, аутор је спровео емпиријско истраживање у

коме су испитаници подвргнути аутентичним тестовним задацима, у оквиру тзв. „task-based“ приступа испитивању језичког знања, захваљујући коме извршење тестовног задатка одражава способност извршења таквог језичког задатка у домену „стварног живота“. Осим тога, испитаници су обучавани да примењују критеријуме за евалуацију перформансе и да преузимају одговорност за испуњавање циљева учења. Резултати истраживања потврђују да излагање студената аутентичним тестовним задацима и методама евалуације и самоевалуације има позитиван утицај на напредак, што је додатно потврђено резултатима оствареним на сумативним проверама знања. (3) Треће истраживачко питање тиче се предзнања и његове укључености у конструкт који је предмет тестирања. Резултати истраживања указују на то да предзнање, односно познавање тематике, има важну улогу у тестирању језика за посебне намене и да позитивно утиче на извршење задатка, чак и код слабијих ученика, тиме што им помаже да испуне циљеве задатка упркос грешкама које се јављају услед слабијег познавања страног језика. (4) Један од циљева ове студије је да истражи да ли аутентични облици тестирања врше позитиван утицај на способност студената да примете сопствени напредак. Резултати спроведеног истраживања указују на то да ставови студената према начину оцењивања игра важну улогу у начину на који се студенти посвећују извршењу задатка, и, сходно томе, утиче на њихов напредак. Такође, резултати истраживања указују на везу између обучавања студената да оцењују сопствени и напредак вршњака и њиховог општег успеха у формативним и сумативним проверама знања. (5) Истраживању спроведеном током израде ове докторске дисертације претходила је анализа потреба која је указала на то да када је у питању енглески језик, постоји неслагање између вештина које свршени студенти поседују и вештина које послодавци на тржишту рада захтевају. Резултати истраживања указују на то да уколико се студенти континуирано излажу аутентичним језичким задацима, а затим и подвргавају аутентичним облицима провере знања, њихове језичке вештине достижу ниво који задовољава потребе тржишта рада. Студија представљена у овој дисертацији на више начина представља допринос теорији и пракси провере језичког знања. Прво, теоријски оквир изложен у првом делу рада доприноси бољем разумевању тестирања вештине говора на енглеском језику. Друго, студија заступа становиште да процес израде језичких тестова треба да узме у обзир ситуациону и интеракцијску аутентичност задатака којима се проверава познавање и

употреба страног језика. Треће, студија предлаже употребу метода којима се минимизира јаз између знања које се стиче током студија и потреба које се јављају на тржишту рада по свршетку студија. Четврто, студија пружа методолошки допринос анализи и изради тестовних задатака. Пето, студија је релевантна у педагошком смислу пошто њени закључци иду у прилог учењу и тестирању које у први план стављају студента. Најзад, будући да је истраживање спроведено у сарадњи са Економским факултетом Универзитета у Крагујевцу, студија доноси бројне предлоге који могу да допринесу развоју курикулума на овој институцији високог образовања.

Кључне речи: провера знања, аутентичност, енглески као језик струке, задатак, пословни енглески језик, конструкт, циљна употреба језика, провера вештине говора.

Научна област: лингвистика

Ужа научна област: примењена лингвистика, тестирање

УДК број: _____

Table of Contents

1 Introduction	1
1.1 Motivation of the study	1
1.2 Research rationale	3
1.2.1 Research questions	6
1.3 Hypotheses	7
1.4 Significance of the study	8
2 Communicative language ability	11
2.1 Communicative and specific purpose language ability.....	11
2.1.1 Language knowledge.....	15
2.1.2 Strategic competence.....	17
2.1.3 Background knowledge	19
2.2 Speaking in the context of CLA	20
2.2.1 What is special about spoken language in communicative settings?	20
2.2.2.1 Characteristics of spoken production/interaction.....	22
2.2.2.1.1 Planning.....	22
2.2.2.1.2 Monitoring and revising	24
3 Testing	25
3.1 Testing communicative language ability.....	25
3.1.1 Communicative language testing	25
3.1.2 Specific purpose language testing	26
3.1.3 Task-based (performance-based) language testing	27
3.2 Types of tests based on the intended use of test results	28

3.2.1 Criterion-referenced tests (CRTs) vs. norm-referenced tests (NRTs).....	30
3.3 Test usefulness: qualities of a language test.....	31
3.3.1 Reliability	33
3.3.2 Validity	35
3.3.2.1 Content-related validity	37
3.3.2.2 Criterion-related validity.....	37
3.3.2.3 Construct-related validity.....	38
3.3.3 Practicality	39
3.3.4 Authenticity	40
3.3.5 Impact	41
3.4 Assessing speaking skills	42
3.4.1 Speaking skills.....	42
3.4.2 Speaking tasks	44
3.4.3 Speaking test task types.....	49
3.4.3.1 Structure: structured, open-ended, role-play.....	49
3.4.3.2 Type of performance: imitative, intensive, responsive, interactive and extensive	51
4 Authenticity in language assessment.....	54
4.1 What is authenticity?.....	54
4.2 Situational vs. interactional authenticity	56
4.3 Critical elements of authentic assessments	58
5 ESP target language speaking tasks and test tasks	63
5.1. Target language use domains and target language use situations	63
5.1.1 Describing a target language use domain.....	63

5.1.2 Target language use domain vs. target language use situations vs. communicative language goals.....	63
5.1.2.1 Identifying target language use tasks.....	64
5.1.3 Construct definition.....	65
5.1.3.1 What to test? Constructs.....	66
5.1.3.2 How to define a test construct?.....	67
5.1.3.3 Construct components.....	68
5.2 ESP speaking tasks.....	70
5.2.1 TLU task characteristics framework.....	71
5.2.1.1 Characteristics of the rubric.....	75
5.2.1.2 Characteristics of the input.....	77
5.2.1.3 Characteristics of the expected response.....	79
5.2.1.4 Characteristics of the interaction between input and response.....	81
5.2.1.5 Characteristics of the assessment.....	84
5.3 Operationalization: developing test specifications and test task specifications.....	86
5.3.1 Test specifications.....	86
5.3.1.1 Test specifications models.....	90
5.3.1.1.1 Alderson, Clapham and Wall (1995) Model.....	91
5.3.1.1.2 Bachman and Palmer (1996) Model.....	92
5.3.1.1.3 Davidson and Lynch Model (2002).....	93
5.3.1.1.4 Douglas's Model.....	94
5.3.1.1.5The CEFR Model.....	94
5.3.2 Test task specifications.....	96
5.3.3 Scoring method.....	97

5.3.3.1 Rating scales	100
5.3.3.2 The CEFR Scales	105
6 Research methodology	108
6.1 Research questions	109
6.2 Hypotheses and expected results	109
6.2.1 Hypotheses	109
6.2.2 Expected results	111
6.3 Data collection and instruments	112
7 Phase 1	115
7.1 From target language use to test tasks	115
7.1.1 Identifying the target language use domain	115
7.1.2 Identifying target language use situations and special purpose speaking tasks	116
7.1.2.1 Grounded ethnography	116
7.1.2.2 Context-based research	118
7.1.2.3 Subject specialist informant procedures	118
7.2 Participants in Phase 1	120
7.3 Research instruments	121
7.3.1 Description of a target language use situation	122
7.3.1.1 Context-based survey	123
7.3.1.2 From general context to specific tasks	128
7.4 Relating TLU speaking tasks to speaking test tasks	132
7.4.1 TLU task characteristics	132
7.4.1.1 Group speaking task (presentation) – TLU task characteristics	133

7.4.1.1.1 The rubric	133
7.4.1.1.2 The input	134
7.4.1.1.3 The expected response	136
7.4.1.1.4 The interaction between the input and response	137
7.4.1.1.5 The assessment.....	138
7.4.1.2 Individual speaking task (short talk/ presentation) – TLU task characteristics.....	139
7.4.1.2.1 The rubric	139
7.4.1.2.2 The input	140
7.4.1.2.3 The expected response	142
7.4.1.2.4 The interaction between the input and response	144
7.4.1.2.5 The assessment.....	145
7.4.2 Test task characteristics	146
7.4.2.1 Group speaking task.....	146
7.4.2.1.1 The rubric	146
7.4.2.1.2 The input	148
7.4.2.1.3 The expected response	150
7.4.2.1.4 The interaction between the input and expected response	152
7.4.2.1.5 The assessment.....	152
7.4.2.2 Individual speaking task	155
7.4.2.2.1 The rubric	155
7.4.2.2.2 The input	156
7.4.2.2.3 The expected response	158
7.4.2.2.4 The interaction between the input and expected response	159

7.4.2.2.5 The assessment	160
7.4.3 Test task specifications	162
7.4.3.1 Group speaking task – test task specifications	163
7.4.3.1.1 The purpose	163
7.4.3.1.2 Construct definition	164
7.4.3.1.3 Learning outcomes	164
7.4.3.1.4 The characteristics of the setting of the test task	165
7.4.3.1.5 Scoring method	167
7.4.3.1.6 Plan for evaluating test usefulness qualities	167
7.4.3.2 Individual speaking task – test task specifications	168
7.4.3.2.1 The purpose	168
7.4.3.2.2 Construct definition	168
7.4.3.2.3 Learning objectives	168
7.4.3.2.4 The characteristics of the setting of the test task	169
7.4.3.2.4 Scoring method	171
7.4.3.2.5 Plan for evaluating test usefulness qualities	171
8 Phase 2	173
8.1 Participants in Phase 2	173
8.2 Research instruments	174
8.2.1 Placement test	174
8.2.1.1 Placement test – participants	176
8.2.1.2 Placement test structure and time allotment	176
8.2.1.3 Task types	176
8.2.1.4 Quick Placement Test administration	178

8.2.1.5 Quick placement test – the analysis of test usefulness qualities	179
8.2.1.5.1 Practicality.....	179
8.2.1.5.2 Reliability	179
8.2.1.5.3 Validity.....	180
8.2.1.5.4 Authenticity.....	180
8.2.2 Task-based approach – authentic speaking tasks	181
8.2.2.1 Differences among experimental groups- task format.....	182
8.2.2.2 Differences among experimental groups - evaluation criteria and feedback	183
8.2.3 “Can-do” checklists (survey).....	184
8.2.4 End-of-semester group oral presentation task	187
8.2.4.1 Group presentation task – assessment.....	188
8.2.4.1.1 Instructor ratings	188
8.2.4.1.2 Peer-ratings.....	188
8.2.5 Final oral exam	190
8.2.5.1 Final oral exam task specifications	190
8.2.5.1.1 The purpose	190
8.2.5.1.2 Construct definition.....	190
8.2.5.1.3 Learning outcomes	191
8.2.5.1.4 The characteristics of the setting of the test task.....	191
8.2.5.1.5 Instructions for responding to the task	192
8.2.5.1.6 Scoring method	192
8.2.5.1.7 Plan for evaluating test usefulness qualities.....	193
8.2.6 Student perceptions survey.....	193

9 Testing hypotheses.....	196
9.1 Hypothesis 1	196
9.2 Hypothesis 2.....	203
9.3 Hypothesis 3.....	208
9.4 Hypothesis 4.....	215
9.5 Hypothesis 5.....	220
9.6 Hypothesis 6.....	226
10 Conclusion	233
10.1 Introduction.....	233
10.2 Summary of main findings.....	233
10.2.1 Using TLU tasks as a model for classroom test tasks	234
10.2.1.1 Deliverable 1 – TLU speaking task characteristics.....	236
10.2.1.2 Deliverable 2 – A desirable CEFR level for spoken interaction/production in TLU context.....	238
10.2.2 Do authentic forms of assessment exert a positive influence on students’ progress?.....	238
10.2.5 Do business students possess the language skills matching the needs of the labor market?.....	248
10.3 Evaluation of the study.....	251
10.3.1 Contributions	251
10.3.2 Limitations and suggestions for future research.....	252
Bibliography	255
Appendices	264
Appendix A: Quick placement test Version 1.....	265

Appendix B: Quick Placement Test Version 2	275
Appendix C: Key to Quick placement test Versions 1 and 2.....	285
Appendix D: Information statement and Consent form (in English)	286
Appendix E: Information statement and Consent form (in Serbian)	288
Appendix F: Data Contribution and Consent form_Companies (in Serbian)	289
Appendix G: Data Contribution Consent form_Companies (in English)	291
Appendix H: Context-based questionnaire: General Context	293
Appendix I: Context-based questionnaire: Business Presentations	294
Appendix J: Assessor’s rating scale (analytic): Rating scale: Oral presentation (group work)	296
Appendix K: Assessor’s rating scale (holistic): Rating scale: Oral presentation (group work)	297
Appendix L: Student self- / peer-assessment rating scale (analytic): Rating scale: Oral presentation (group work)	298
Appendix M: Student self- / peer-assessment rating scale (holistic): Rating scale: Oral presentation (group work)	299
Appendix N: Assessor’s (and student self/ peer-rating) rating scale (holistic): Rating scale: Individual short presentation.....	300
Appendix O: Self-evaluation checklist - shuffled (spoken interaction and production in English)	301
Appendix P: Self-evaluation checklist – ordered (with corresponding CEFR levels)	304
Appendix Q: Can-do evaluation checklist - shuffled (spoken interaction and production in English for subject specialist informants).....	307
A:.....	307
Appendix R: Self-evaluation checklist – ordered (spoken interaction and production in English for subject specialist informants)	311

Appendix S: Student attitudes questionnaire (A in English, B – in Serbian).....	314
Appendix T: Self-evaluation checklist - shuffled (target; with or without help).....	316
About the author (biography)	320

1 Introduction

This chapter offers a brief introduction to the principal aspects of this doctoral dissertation. The author's motivation to conduct the study and research rationale are presented. This is followed by an outline of the research questions and hypotheses. Finally, the chapter ends by stating the intended significance of the research.

1.1 Motivation of the study

As the abstract of the thesis indicates, the research investigates authentic forms of assessment in the context of testing English for specific purposes speaking skills. By using grounded-ethnography and context-based research techniques, the study explores the real life domain pertaining to the use of Business English for business communication in Serbian companies. To have a better understanding of specific purpose language tasks, the study collaborates with subject specialist informants who feed the research with specific characteristics of the context and tasks taking place in work settings. The obtained information is then analyzed by the means of the Task characteristics framework, resulting in a set of target language use task characteristics, based on which authentic speaking test tasks are developed. These tasks are then applied in the educational domain and their effects on learners' perceptions and progress are observed and investigated.

The present study is motivated mainly by two factors: the author's EFL teaching career and the mismatch that exists between the academia and the real life needs when it comes to English language. The first factor that has sparked this study is the author's experience as an English language instructor in the context of higher education in Serbia. Although all universities promote the idea of teaching and learning English language, the methods and the settings in which students learn this language are often constrained by practical considerations: time, space, and the available personnel. Speaking of time, the author refers to the number of contact hours per week dedicated to studying English within a particular study program. The considerations of space and personnel are linked together as they refer to accommodating the language needs of fairly large groups of students by two or three instructors employed at a given faculty. In such circumstances, many instructors struggle with maintaining the quality of instruction, while, at the

same time, they are required to assume the role of test developers. Apparently, the majority of assessment practices taking place in university settings refer to summative assessment, resulting in a midterm or a final grade. While giving students grades comes as a natural outcome of the teaching process, it seems that there is little space provided for alternative assessment methods - those promoting independent and collaborative learning- with students taking the accountability for the actual learning outcomes. In the same vein, regardless of the fact that many curricula take the approach to teaching skills, the approach to testing is quite often restricted to assessing grammatical and the knowledge of vocabulary by the means of multiple-choice testing format. Not necessarily underestimating the reliability of such testing practices, the author questions their authenticity, as well as the validity of test scores and the inferences based on them testifying that test takers have the ability to actually speak the language.

Another factor inspiring this research is related to the apparent mismatch between the real life language needs and the learning outcomes envisaged by university curricula. Putting the author's intuition aside, his experience in conducting in-house English language trainings for middle and senior management in "Zastava Upholstery" company indicated that managers with business background had very limited oral English language skills. However, their topical knowledge as well as the use of specific purpose vocabulary were quite satisfactory. This pointed out the issue of business students not getting enough language practice in performing real life tasks throughout the course of their studies. In addition to this personal experience, some survey results (discussed in the following chapter) indicate that there is a growing demand for skilled labor force, capable of actively using English language for business communication. In the same vein, Green (2014) emphasizes that educators need to be aware that there has been a shift in the focus of language education:

...The older ideal of language education was for learners to develop an appreciation of the finest in the foreign literature and culture. This aim has gradually been displaced in many Ministries of Education and other policy-making bodies by the more utilitarian view that knowledge of foreign languages is a basic skill that has economic value: readying workers to participate in international markets. Learners themselves often wish to acquire a language not so much to access the cultural highlights as to help them to travel, to build a new life in a foreign country, to do business...(p. 175)

Considering that there have been many changes in the economic life in Serbia, the author believes that this study will point out the curricular changes that need to bridge the gap between

learning and testing English for academic purposes and preparing students for solving real life tasks. The research rationale that follows aims at pointing out the significance of the present study, by placing it in the context of the professional domain in Central Serbia.

1.2 Research rationale

A TEMPUS project named “*Reforming Foreign Language Studies in Serbia*” was implemented with the purpose of modernizing the manner in which foreign languages are taught, studied and assessed in order to bridge the gap between academia and the real needs outside university settings. One of the project strands was dedicated to working closely with labor market representatives in order to determine which foreign language skills are deemed desirable for prospective employees. At the same time, the project aimed at facilitating curricular reforms that would meet demands for highly skilled professionals in the work settings. One of the project deliverables, resulting from a comprehensive survey, was a study published under the title *Philology Studies and Labor Market Needs*, which indicated that most enterprises, participating in the survey, expected their employees to actively use at least one foreign language – predominantly English (REFLESS, 2012:42). The survey was conducted in collaboration with Serbian Chambers of Commerce, and included a representative sample of respondents, mainly from the private sector (86%). A subsequent market needs analysis showed that employers expected their employees to be able to orally communicate in English, given that their overall English language competence was perceived as their ability to speak this language, all leading to the conclusion that employees’ verbal skills are deemed as more important than any other language skills (REFLESS 2012:43).

A study, entitled *The Evaluation of Studies and Professional Success of Graduate Students in Serbia and the Region*, published in 2014 within the CONGRAD TEMPUS project, indicates that more than half of the university graduates seek employment in the private sector (51%). Additionally, it indicates that, in the majority of cases (71.7%), job posts require that graduates perform tasks based on the skills and knowledge gained during their university studies (CONGRAD, 2014:9). If we compare that to the survey results collected within the TEMPUS REFLESS project, mentioned above, it becomes clear that enterprises in Serbia are mainly privately owned, and it is the private sector where students are likely to seek employment. In line with that, higher education institutions are facing the task of meeting a growing market demand

for qualified and highly educated individuals capable of performing real life tasks. The latter study indicates that 62% of graduate students were required to apply exactly the same knowledge and skills they acquired in the course of their studies, emphasizing the need for the curriculum to be relevant to the settings outside university (p. 11). On the other hand, a certain number of graduate students responding to the survey claim that curricula are often impractical and obsolete, leading to the conclusion that university administrators should identify and modernize such curricula in order to make them relevant to the real life domain.

The analysis of the aforementioned studies points out the following indicators of changes in the economic forum of Serbia: privatization, foreign language knowledge requirements, job-seeking strategies, the role of the National Employment Service, the language of job titles and job advertisements, the prominent role of English language. In the circumstances of transitional economy where public companies transform into privately owned ones, as well as in the business environment characterized by direct foreign investments, many companies opt for hiring professionals who do not only possess field-specific knowledge, but who can also communicate in foreign languages, English in particular. Judging by the research results, the role that English language plays seems so important that employers and HR services consider the ability to communicate in English as one of the job requirements, which is best evidenced in advertisements on one of the most visited websites for prospective employees (www.poslovi.infostud.com). Given the importance that communications skills are given, a conclusion can be drawn that the communicative language model plays a crucial role in equipping students with the skills they need in their future career.

Changes in economy have affected job-seeking strategies employed by prospective applicants. According to the research conducted within CONGRAD TEMPUS project, prospective applicants apply the following job-seeking strategies:

- seek employment through social networks and relatives (32.8%),
- browse websites looking for online job advertisements (21.2%),
- address the National Employment Service for help with employment (12.7%).

(CONGRAD, 2014: 8).

The last two findings indicate an important change in the role of the National Employment Service (NES) as the main mediator between employers and prospective applicants. Namely, in the period prior to the start of privatization of the public sector, the NES played the most important role in helping applicants find employment. There were two reasons why this was the case. First, every individual had (and most likely still has) their file open with the National Employment Service, containing data related to their educational background, employment history and personal information. This fact implies that the National Employment Service possesses the largest database of prospective employees in the country. Second, there was a tradition for every company (prior to the privatization, they were all public) to hire employees through the NES, whose role was to perform selections and facilitate the hiring process. Consequently, the National Employment Service served as a large database of job advertisements, given that it cooperated directly with prospective employers. The process of privatization and the Internet introduced significant changes to the role that the NES had had prior to it:

- private companies offer direct employment, facilitated by their own HR departments;
- specialized employment agencies provide employment mediation services; and
- job advertisements are published on specialized websites.

The language of job advertisements is another indicator that changes have taken place in the work environment in Serbia. Research papers dedicated to analyzing Anglicisms in advertisements, published after 2008, point out that when it comes to job titles, 46.66% of them are derived from the English language. In addition, many job titles are used in their “raw” or original English form, whereas the rest of the text in advertisements is published in Serbian (Milanović & Milanović, 2012, and Milanović & Milanović, 2012a, and 2012b). The same research indicates that more than 30% of job advertisements are published in a foreign language, prevalently in English, which implies that applicants are required to submit their CVs and job applications in the language of the advertisement. Consequently, the prospective applicants need to possess sufficient language knowledge and communication skills to compose the cover letter and their curriculum vitae in English. Additionally, if the text of a job advertisement is published in English, and if the required documents are in English, it comes as an unwritten rule that shortlisted candidates will be interviewed in English. The authors of the *Studies of Philology and the Labor Market Demands Study* conducted a survey whose results indicate that employers and

their HR officers interview and often “test candidates assessing their communicative ability” in English (REFLESS, 2012: 43). This implies that job applicants are expected to have mastered English prior to being employed.

In summary to this chapter, it should be noted that the needs analysis conducted for the purpose of the research relevant to this doctoral thesis indicates the following:

- there are research projects and studies indicating that there is a gap between the skills and knowledge that graduate students gain in the course of their higher education and the skills and knowledge that they are expected to demonstrate in work settings;
- English language (especially oral skills) is highly valued and considered as an indicator of an overall communicative ability in this language;
- companies that are performing business operations at the territory of Serbia are mainly privately-owned;
- prospective employers often publish job advertisements online; many of the advertisements are published in English (about 30%) and require that employees be able to actively use it.

1.2.1 Research questions

In line with the findings discussed above, this doctoral thesis aims at providing answers to the following research questions:

- 1) Can target language use situation tasks be used as a model for authentic classroom test tasks?
- 2) Do authentic forms of assessment exert a positive influence on students’ progress?
- 3) Should background knowledge be tested in specific purpose speaking assessments?
- 4) Do authentic forms of assessment exert a positive influence on students’ awareness of their own progress?
- 5) Do business students possess the language skills matching the needs of the labor market?

In accordance with the research questions stated above, this thesis will be based on a research investigating the assessment of spoken skills in English by the means of employing authentic test tasks. The research will be conducted in two phases:

Phase 1 – collecting data in collaboration with 25 subject specialist informants representing the real life domain (labor market at the territory of the Municipality of Kragujevac); and

Phase 2 – collecting data in the domain of higher education, on the sample of 150 business students enrolled in the Faculty of Economics (modules: *Management, Accounting and Business Finance*, and *Marketing*), University of Kragujevac.

The data collected during the two phases of the research will be analyzed and used to test and validate the hypotheses presented in the following chapter.

1.3 Hypotheses

The research conducted for the purposes of this doctoral thesis aims at investigating spoken English language skills assessed through formative and summative test methods, by the means of authentic input material and test tasks. The test tasks used in the research come as a product of a thorough analysis of target language use situations in which language users complete various real life language tasks (Bachman and Palmer, 1996). In this way, the author will investigate authenticity of test tasks that are created based on the TLU situation analysis, as well as the effect that authentic speaking tasks have on students' progress. Bearing in mind that class assessment within any particular curriculum has two purposes – to check both student progress and attainment of learning objectives, and to ensure that future employers' expectations are met – the research aims at determining the extent to which authentic test tasks may have a formative role in facilitating students' progress.

Based on the theoretical framework presented in the first part of the dissertation, an empirical research will be conducted with the purpose of testing and validating the following:

H1: The examinees who have been thoroughly trained to apply evaluation criteria demonstrate a better overall performance in the final oral exam in comparison to the examinees who have not been thoroughly trained on applying analytic and holistic scoring criteria in assessing their own and the performance of their peers.

H2: Performing on a task requiring that test takers should possess background knowledge related to the field of *Marketing*, the Control group demonstrates very similar results to the more successful of the two experimental groups.

H3: End of semester survey results indicate that more than two thirds of the examinees demonstrate positive perceptions of authentic tasks, as well as of the system of evaluation and self-evaluation that they have been exposed to.

H4: End of semester self-evaluation questionnaire results indicate that at least 70% of the Control group's responses provided to estimate their target skills match the responses provided at the beginning of the semester.

H5: End-of-semester self-evaluation results indicate that at least half of the sample in the Experimental groups achieved progress by one CEFR level, as corroborated by the Second placement test results.

H6: The highest agreement in responses to the "Can-do" survey is the one between subject specialist informants and Group 1 subjects.

If the author's assumption that authentic test tasks and test performance evaluation methods correlate with target language use tasks and methods of evaluation is proved to be true, the conclusion to be drawn is that such forms of assessment play a formative role bringing students' language skills closer to the requirements of the labor market. Employers have certain expectations of the language skills their prospective employees should possess before they join the company, so it is university where these skills need to be developed.

1.4 Significance of the study

By investigating authentic forms of assessment in the context of testing ESP speaking skills, the research has important implications for a range of areas.

First, this study has a theoretical significance in that that it not only contributes to a better understanding of ESP speaking assessment, but it offers practical solutions to enhancing the authenticity of assessment endeavors. Through the application of the task-based approach to testing oral English language skills in the context of ESP language learning, the study does not

aim at undermining the so-called pedagogical tasks, but draws the educators' attention to careful consideration of test constructs and alternative assessment practices.

Second, this study helps contribute to a deeper understanding of situational and interactional authenticity, equipping prospective test developers with understanding of critical elements pertaining to authentic assessments. The consideration of what constitutes authentic assessments helps the test developers claim that their assessments have the real life value, i.e. the value outside testing contexts.

Third, this study advocates bridging the gap between academia and industry by providing theoretical foundations and practical tools aimed at fostering collaboration between developers of specific purpose language tests and informants from the real life domain.

Fourth, the study has a methodological significance for the ESP testing field in that it offers a tool for ensuring the comparability and correspondence between target language use tasks and test tasks. The Task characteristics framework presented in this study may be of significance to language testers who strive for enhancing the authenticity of the assessment process.

Fifth, the study has a pedagogical significance. By examining students' perceptions, it investigates the influence that authentic assessment methods exert on student learning and progress. At the same time, it emphasizes the importance of collaborative, independent and student-centered learning through the application of formative assessment methods.

Sixth, the findings from this study can contribute significantly to the curricular changes at the host institution – the Faculty of Economics, University of Kragujevac. The research results aim at pointing out strengths and weaknesses of the English language 2 course syllabus. If the research results show that authentic test tasks and evaluation methods exert a positive influence on students' progress and that they stimulate learning, the assessment practices and the course syllabus may benefit from the research deliverables – speaking test task specifications and the plan for evaluating test usefulness.

The following chapters offer theoretical foundations for the research (Chapters 2-5), discussing the following topics: communicative language ability, testing, authenticity, and ESP target language tasks and test tasks. Chapters 6 – 9 present the actual research, outlining its

stages, methodology underlying the use of research instruments and results. Finally, Chapter 10 concludes the thesis by outlining the main findings and offering a critical perspective of the study's contributions and limitations, accompanied with suggestions for future research.

2 Communicative language ability

Special purpose language testing is considered to be a variety of communicative language testing (Douglas, 2000), which developed under the influence of communicative language ability theory in 1980s and 1990s. To have a clear understanding of the principles on which special purpose language testing is based, this section will outline communicative language ability theories and the communicative testing model as a foundation for special purpose language testing model.

2.1 Communicative and specific purpose language ability

Communicative competence as a term dates back to 1970s, when Hymes (1972) proposed that in addition to language knowledge, individuals' use of the knowledge to perform tasks in real-life situation must also be taken into account. These tasks require social interaction and take place in a particular context, each influencing the communication that takes place in a given moment. The sociolinguistic component to the study of L1 that Hymes added in his works in 1972 and 1974 influenced the work of Canale and Swain in 1980. They built on Hymes' ideas in their attempt to design a framework that will facilitate the design of curricula and English as a Second/Foreign Language test development projects. This framework describes communicative competence as an ongoing interaction among *grammatical competence*, *sociolinguistic competence* and *strategic competence*. In other words, communicative competence was seen as a dynamic process which draws upon an individual's knowledge of grammatical rules, socio-cultural norms of the world in which an individual operates and strategies for handling "breakdowns in communication" (Canale and Swain 1980 in Young, 2008:97). Revisiting the model in 1983, Canale added another competence to the model (discourse competence) justifying it by the requirements of cohesion and coherence in language production (in Weir, 1993:8).

The work of Hymes in the 1970s and that of Canale and Swain in the 1980s influenced further development of communicative language model. Defining *communicative language ability* (CLA), Bachman says, "CLA can be described as consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use" (1990: 84). Bachman created the framework of

CLA, proposing that it should contain the following components: *language competence*, *strategic competence*, and *psychophysiological mechanisms* (Figure 2.1). The language part of the model, often referred to as language competence, involves a “set of specific knowledge components” that are engaged in the process of communication, hence the stress on the communicative language use (ibid.). In addition, the CLA model includes a set of metacognitive strategies, also known as strategic competence that allows for analyzing the context and employing context-appropriate strategies enabling individuals to participate in communication.

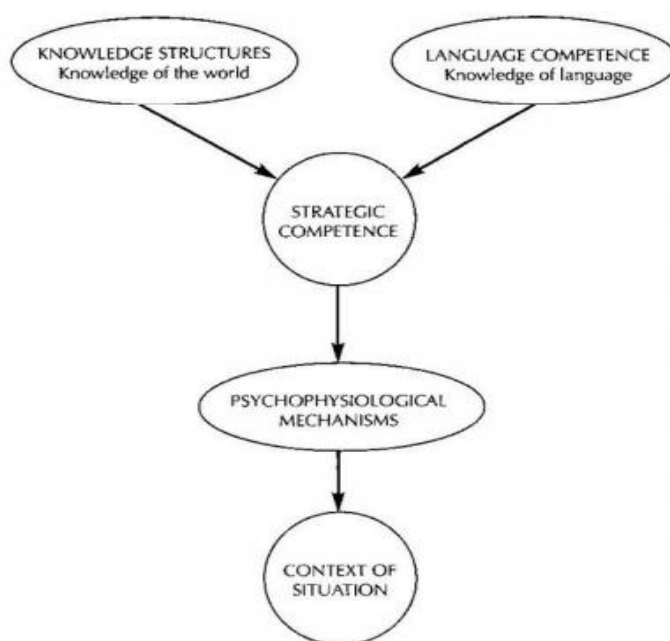


Figure 2.1 Components of communicative language ability in communicative language use
(Bachman, 1990:85)

Psychophysiological mechanisms refer to “the neurological and physiological processes [...] that are employed during execution phase of language use (Faerch and Kasper, 1983 in Bachman, 1990:107). This model was reworked a couple of years later, when Bachman and Palmer (1996), building on the model from 1990, proposed a five-componential model of communicative language ability consisting of the following: *language knowledge*, *topical (background) knowledge*, *personal characteristics*, *strategic competence*, and *affective factors* (Figure 2.2). This model brought the idea of communicative ability as a dynamic process that does not reside solely in an individual, but is influenced and directed by a number of internal and

external factors within a certain context. As in earlier works on communicative competence, this model sees the communicative language ability as an interactional process.

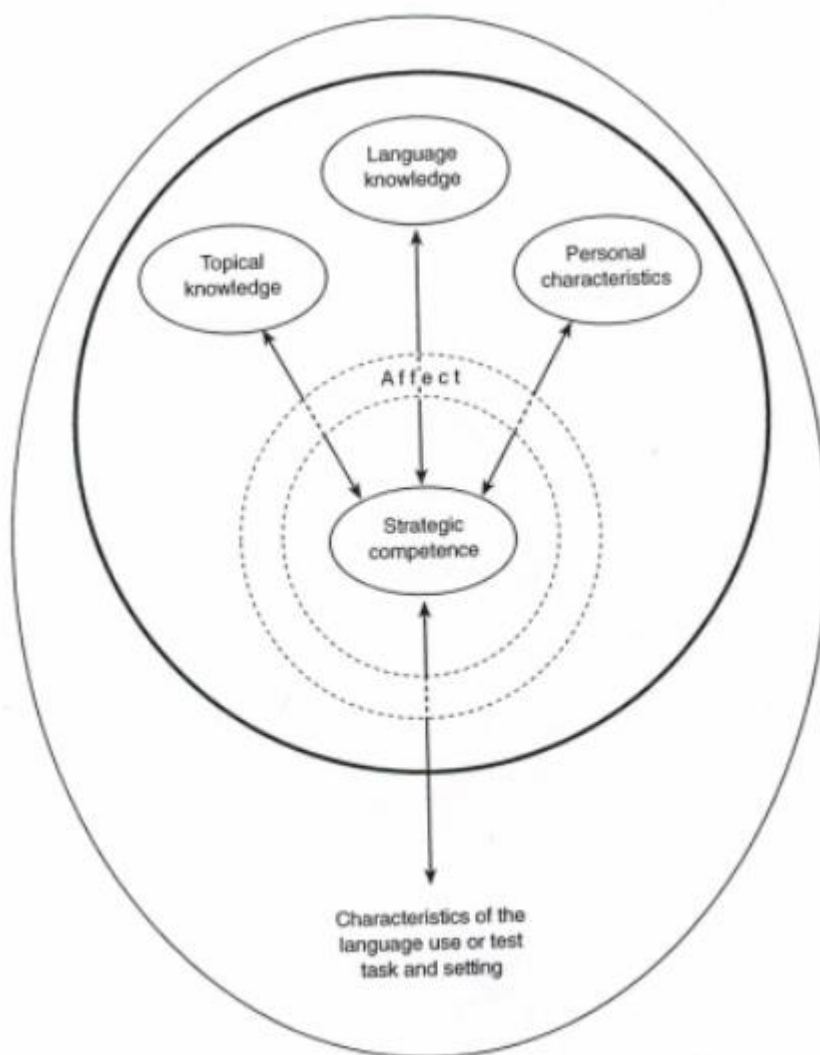


Figure 2.2 Components of Communicative language ability (Bachman and Palmer, 1996: 63)

At about the same time, Celce-Murcia et al. (1995) proposed a componential model of communicative competence that included *discourse competence*, *linguistic competence*, *actional competence*, *socio-cultural competence*, and *strategic competence*. Actional competence is a component that refers to performing language tasks resulting in an interaction, therefore Celce-Murcia revisited this model in 2007 renaming actional into *interactional competence*, and added additional competence that takes fixed expressions and phrases into consideration, naming it

formulaic competence. The six-component model of communicative competence envisages a constant interaction of the aforementioned competences within a particular language context (see Figure 2.3).

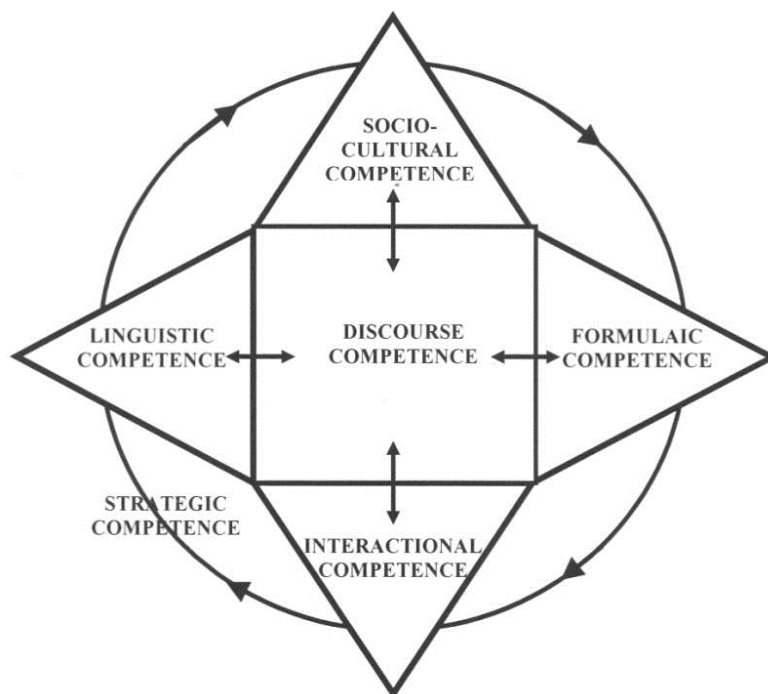


Figure 2.3 Six-component scheme of communicative competence (Celce-Murcia, 2007:45)

Specific purpose language ability

In his attempt to understand what constitutes a construct of *specific purpose language ability*, Douglas draws upon Bachman and Palmer’s notion of communicative language ability and understanding of an external context in which language learning and communication take place. He also builds on Chapelle’s *interactionist view* of construct definition, by which characteristics of test takers (including their language knowledge and strategic competence) interact with characteristics of context resulting in both sets of characteristics being affected (1998 in Douglas, 2000:24). One of the main results of this interaction, according to Chapelle, refers to the limitation of linguistic choices imposed by a specific context. In other words, the external context is “a major factor in the engagement of specific purpose language ability” (p.25), which occurs as a result of the interaction between language ability and specific purpose

background knowledge by the means of strategic competence (Figure 2.4). Specific contexts and target language domains will be discussed in more detail in Chapter 4.

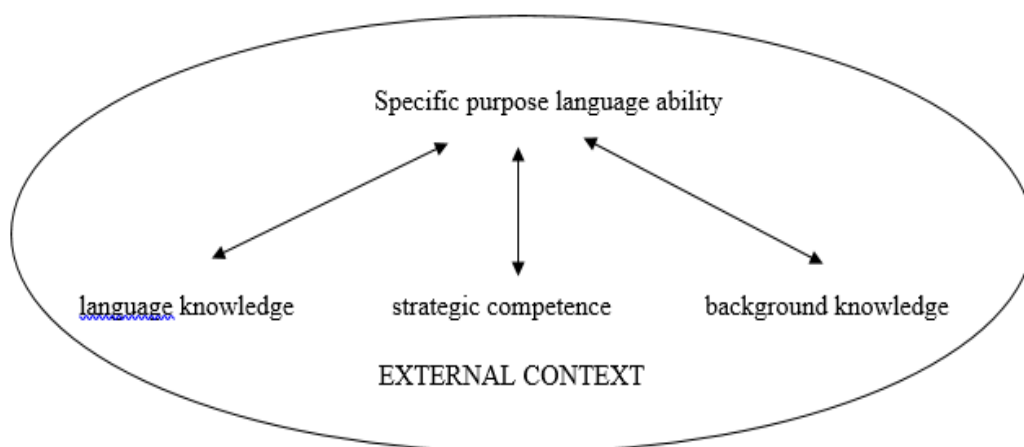


Figure 2.4 Specific purpose language ability

It should be noted that in this thesis, communicative language ability model is discussed in terms of its applicability in communicative language testing, as well as in special purpose language testing, which Douglas regards communicative by definition (2000). In the subsequent chapters, we will discuss language knowledge and strategic competence as constituent parts of CAL; in addition, we will define background knowledge and its role in special purpose language testing.

2.1.1 Language knowledge

Language knowledge can be defined as a “domain of information in memory that is available for use by the metacognitive strategies in creating and interpreting discourse in language use” (Bachman and Palmer, 1996:67). To test language knowledge, it is important for test developers to know what it includes, though, it should be noted that there are different classifications of language knowledge (or language competence) in literature on language assessment. Bachman, for example, groups morphology, syntax, vocabulary, phonology/graphology under the component of *grammatical competence*, while cohesion and rhetorical organization are grouped under *textual competence*; both grammatical and textual competence are elements of **organizational competence** category. Organizational competence

can be regarded as a set of abilities that are employed in structuring grammatically correct elements and combining them appropriately so that they form a written or spoken text. The other category is **pragmatic competence**, consisting of *illocutionary competence* which itself is a set of various functions (ideational, manipulative, heuristic, and imaginative function); and sociolinguistic competence with their constituent elements necessary to analyze the socio-cultural and discursal features of a context (sensitivity to dialects, register, and nature; imaginative function, cultural references and figures of speech). Bachman discusses the components of language competence arguing that language testers never include all of them in a single test, but, nevertheless, they should be aware of what constitutes this competence (1990: 87, see Figure 2.5). Later on, Bachman and Palmer gave up such division, offering alternative categorization underlining that language knowledge involves organizational knowledge, grammatical knowledge, textual knowledge, pragmatic knowledge, functional knowledge and sociolinguistic knowledge (Bachman and Palmer, 1996: 66-70).

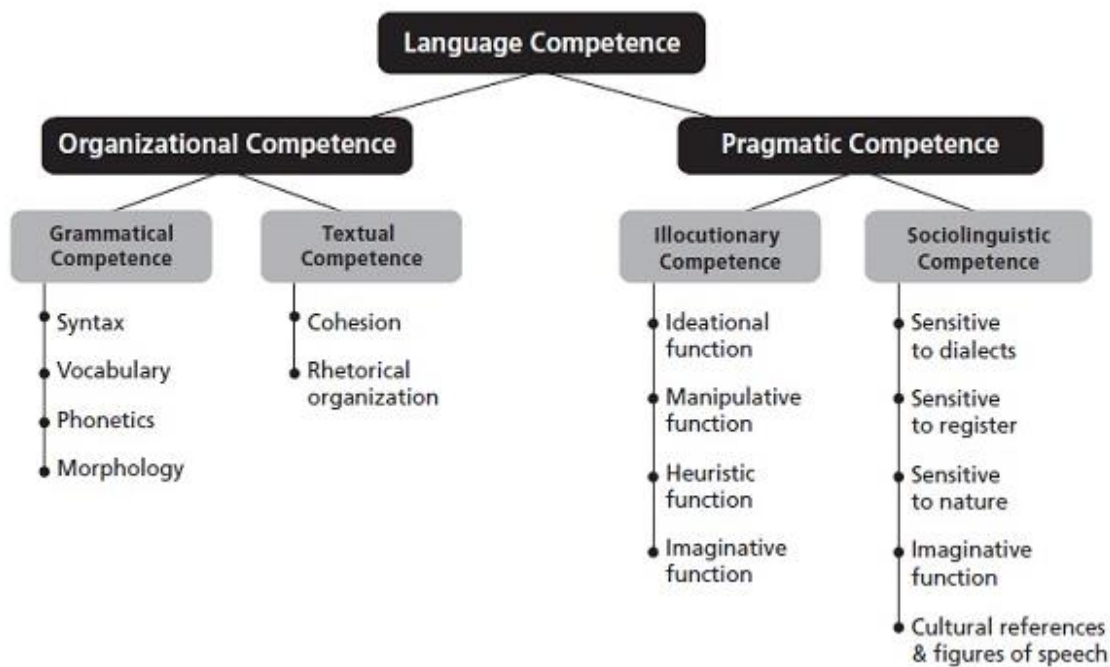


Figure 2.5 Components of language competence (Bachman 1990 in Castillo Losada et al. 2017:90)

In his work dedicated to assessing listening, Buck makes distinction between two types of knowledge: *declarative knowledge* and *procedural knowledge*. The former is related to knowing

facts about something, whereas, the latter refers to the knowledge of procedures for putting knowledge into action. Hence, declarative knowledge is of limited use unless it is combined with procedural knowledge for fulfilling a particular purpose (Buck, 2001:14). Weigle, discussing writing assessment and building on Grabbe and Kaplan's modified taxonomy, states that language knowledge can be divided into three broad categories: linguistic knowledge, discourse knowledge, and sociolinguistic knowledge. These broad categories can be further subdivided into smaller constituent components (Grabbe and Kaplan, 1996 in Weigle, 2002:30). The knowledge of grammar and vocabulary are often assessed as the knowledge of structures and discrete points, so many authors question their value in the context of communicative language assessment (Read, 2000:3). Powers believes that communicative competence as a concept involves the ability of learners to integrate various elements, such as lexis, grammar, strategic competence and others in order to achieve their communicative goals (Powers, 2010:2). Discrete point assessment can be justified by the claim that grammar and vocabulary are inextricable components of communicative language use, and as such, they should be assessed as well.

2.1.2 Strategic competence

In communicative language assessment model, strategic competence, as well as language knowledge may be assessed, provided it falls under the construct definition. Defining communicative language ability model, Bachman underlines that communicative language ability can be understood as an interaction among language knowledge, strategic competence involving mental capacity to implement "the components of language competence in contextualized communicative language use", and psychophysiological mechanisms enabling actual execution of language (1996: 84). Castillo Losada et al. define strategic competence as "the ability to compensate in performance for incomplete linguistic resources in a second language" (2017: 90). Building on a speech production model developed by Faerch and Kasper in their attempt to design a psycholinguistic model of speech production (Faerch and Kasper 1983 in Bachman, 1990: 100), Bachman proposes that strategic competence should include three components: assessment, planning, and execution (ibid.). Bachman and Palmer adapted this model of strategic competence in 1996, realizing that in the context of language assessment, as well as in the context of target domain language use, individuals rely on their topical schemata, language schemata, and affective schemata in order to engage in solving language tasks

(Bachman and Palmer, 1996:65-75). The procedure of solving a task means that a speaker should follow a number of steps: set a goal; estimate the task and its constituting components (while estimating their own language knowledge and background knowledge at the same time in order to determine whether they have sufficient knowledge to complete the task); and, finally, make a plan on how to draw on their language knowledge, topical knowledge, and affective schemata in order to tackle the task. There are authors who claim that strategic competence includes the control of linguistic execution (Douglas, 2000:82; Weigle 2002:44), where language user responds to a prompt/task by employing appropriate language and background knowledge “engaging it in either production or comprehension by the means of appropriate psychophysiological mechanisms” (Bachman, 1990 in Douglas, 2000:82). In the context of language assessment, the strategies mentioned above will become an integral part of a test construct, regardless of whether they are explicitly stated or not. In some situations, strategic competence is explicitly stated as a part of construct definition, regardless of whether the construct is defined by following componential or holistic approach. In other words, if the purpose of an assessment is to investigate constituent components of the strategic competence, then construct definition will reflect goal setting, assessment, planning, and execution stages of the strategic competence. To illustrate this, we can take for example computer-based language tests which offer test takers a number of options including: replaying the recording in tests of listening comprehension, word definitions in tests of reading comprehension, navigation through test items enabling test takers to skip items or go back to them, the possibility to change the answer in the case of a mistake, the option to hide/show a clock on the user interface for better time *Management* (the user can decide to hide the clock if they find it distracting). In such cases, test results not only reflect test takers’ language knowledge, but their strategic competence, as well. For this reason, construct definition should delineate components of the strategic competence which are actually being tested (more in Chapelle and Douglas, 2006:12). When it comes to assessing languages for specific purposes, Douglas considers strategic competence to be a link between the external, situational context and internal language knowledge and background knowledge that are engaged in the process of responding to a test task. It is also his view that strategic competence is inherent to all language use situations, outside or within the testing context, but it is the purpose of assessment and intended use of test results that determines whether it will be explicitly stated in the construct definition (Douglas, 2000:38).

2.1.3 Background knowledge

Background knowledge is the type of knowledge which is directly related to the topic, link or stimulus, and its presence in a testing context is usually a reason for dispute among researchers. The main reason for this lies in the threat that background knowledge in a language assessment may contaminate the score due to a construct-irrelevant variance; hence, it is hard to expect consensus as to whether it should be tested or not. The traditionalists' view raises concerns in terms of assessment validity and fairness, since the test takers who have been more familiar with the topic will be more likely to solve language test tasks with more success than those who do not know much about it. In such case, the results do not necessarily reflect test takers' language ability, but also their background knowledge. Communicative language testers, on the other hand, claim that there are three possibilities concerning background knowledge and its presence in the construct definition (Weigle, 2002:45): (a) background knowledge is not included in the construct definition as it may cause fairness and validity issues, giving advantage to certain test takers whereas disadvantaging others. Background knowledge is not included in construct definition when test takers are not expected to possess the same topical knowledge, such as in language programs; academic, professional and vocational training programs, etc.; (b) background knowledge is included in the construct definition when test takers are expected to have more or less similar background knowledge resulting in tests items being related to specific topical knowledge (Douglas, 2000:39). This is often the case in assessing languages for the purposes of employment, selections for vocational programs, language for specific purpose programs, etc.; (c) background knowledge and language ability are defined as separate constructs because test developers do not know whether the group being assessed possess homogenous background knowledge, but test users still require that inferences be made about both their language ability and areas of background knowledge. This often happens in specific purpose language programs, 'where the language is being learned in conjunction with topical knowledge related to specific academic disciplines, professions or vocations' (Bachman and Palmer, 1996: 125); it is also relevant to employment contexts where potential employees are required to use the language while performing their future job-related tasks (ibid.).

2.2 Speaking in the context of CLA

Traditional testing practice divides overall language ability into two broad categories, based on the cognitive process and senses involved in processing and responding to input: receptive and productive. They are further subdivided into two more skills: reading, listening, writing and speaking. The current testing practice, however, observes the skills as complementary in language performance, although they can be assessed following either stand-alone or integrated testing principle. In this chapter, we will discuss speaking ability in the context of Communicative Language Ability approach in order to understand how to assess it so that test scores provide valid inferences of test takers' speaking ability.

2.2.1 What is special about spoken language in communicative settings?

Sound is one of the most distinctive characteristics of spoken language. When they speak, people produce sound which reveals a great deal of information about the very speaker. For example, based on pronunciation, a person's origin, social and educational background can be revealed. Based on the intonation of their utterances, the volume and the pitch, people convey much more than just a message – they demonstrate their feeling and attitudes, etc. In assessment contexts, the sound of speech includes several aspects that are normally included in foreign language curricula, and are therefore taught and tested – individual sounds, pitch, volume, speed, pausing, stress and intonation (Luoma, 2004). The purpose of assessment and the kind of information that is to be obtained through assessment help test developers decide what aspects are relevant to a particular testing purpose.

Spoken grammar is another characteristic of spoken language that refers to grammatical forms and structures that speakers produce and combine correctly, while delivering speech. Ochs (1979) states that speech itself falls into two categories: *planned* and *unplanned*. The former refers to speaking events that have been prepared and rehearsed (for example, lectures, presentations, prepared speeches, expert discussions, etc.), whereas unplanned speech events are product of a moment and a situation, often in the form of a reaction to an external input (for example, an answer to a question, a reaction to somebody's remark, etc.). When planned speech is delivered, the circumstances normally require a higher degree of formality than in unplanned speech, as well as more complex grammar structures, clear and correct pronunciation, and often

special purpose (even technical) vocabulary. Luoma (2004) argues that in assessments spoken grammar should be evaluated by considering the following:

- speech consists of idea units, not sentences,
- spoken grammar tends to be simpler than written grammar,
- pauses and hesitation markers are punctuation in speaking,
- in interactive speaking, constructing an idea is a joint effort, and
- grammar in planned speech is more complex than in unplanned
- planned and unplanned speech differ in levels of formality and choice of vocabulary.

Equally, or even more important than spoken grammar, vocabulary in speech is the basic tool in oral communication. Like grammar, spoken vocabulary has its own peculiarities, and there are certain expectations regarding learner vocabulary and progress that learners make as they move from lower to higher proficiency levels. Considering that the focus of this thesis is the assessment rather than the development of speaking ability, we will discuss some characteristics of spoken vocabulary that test developers should bear in mind. First, it is common to consider it a sign of high level of proficiency when language learners use rich and complex vocabulary correctly. However, Read finds that it is equally important to use common words naturally and correctly, as this is also a sign of proficiency (2000). Second, unlike written language, which, in specific purpose situations, lends itself to the use of specific/technical words, spoken performance includes many generic words, regardless of their lack of specificity. For example, speakers often use demonstratives to refer to persons/objects that are familiar in a given context (either because they can be seen, or because other participants know what the demonstratives refer to). Third, native speakers, when they engage in interactive and informal conversations, often use *vague* words, such as “*thing, whatsit*” when they cannot recall the actual word, or when they expect the interlocutor to complete the missing word (in their mind or by actually saying the word). Fourth, it is natural for speakers to use words and phrases intended to give them time to assess the situation and think of next thing to say. These floor-keeping techniques involve using fillers (e.g. *you know, sort of*) and hesitation markers (e.g. *um, ah*), as well as fixed phrases, which competent speakers use on appropriate occasions (e.g. *How nice of you to say that!*). Fifth, it is common even for proficient speakers to make slips, errors and omissions in speech. These tend to be attributed inflated significance in assessments, affecting test takers’ grades as

assessors often consider them as an attribute of poor knowledge or preparedness. Luoma suggests that test developers consider writing rating rubrics that will provide assessors with the opportunity to reward test takers for using correctly the categories of words discussed above as well as to devise the way in which slips will not be given exaggerated importance, especially in cases when test takers, noticing their own mistake, correct themselves (2004:19).

2.2.2 Spoken production and interaction

Some authors make distinction between spoken production and spoken interaction (Council of Europe, 2001). The former refers to situations when a speaker addresses others through extended speeches, very much like monologues, e.g. by delivering lectures or public speeches. The interaction, on the other hand, tends to be more natural and informal, with the shared responsibility for constructing the spoken exchange. However, Green argues that these two can be better regarded as “the two extremes of a continuum” (Green, 2014: 128), since they cannot be entirely independent in speech. For example, although speaking in public involves a great deal of preparation and rehearsal before delivery, the actual delivery does not exclude exchanges between the speaker and the audience, adding the elements of spoken interaction. Similarly, no matter how spontaneous and informal interaction between participants in a speaking situation is, it does not exclude pauses during which interlocutors prepare for the next exchange.

2.2.2.1 *Characteristics of spoken production/interaction*

This chapter offers an overview of the main characteristics of spoken production and spoken interaction respectively. The author provides a brief overview of *planning* and *monitoring and revising*, offering a brief description of what these characteristics entail.

2.2.2.1.1 *Planning*

When they engage in speaking production or speaking interaction, speakers spend more or less time planning how to construct the message. Spoken production, however, tends to include more careful *planning*, with the speaker spending more time on preparing the utterance or the speech. This results in the extended spoken language sample, which has many characteristics of a monologue, but it also leaves some room for interruptions, usually in terms of the audience or the interlocutor(s) questions and comments. The linguistic components of the

produced speech can be analyzed in terms of grammar and vocabulary used in the utterance. When they have enough time to prepare their speech, addressers use more complex syntax and grammar, more or less skillfully combining coordinate and subordinate clauses. The vocabulary, naturally, varies according to the purpose and context, ranging from general to specific purpose. Students delivering a presentation on a general topic, one that does not require too much preparation and specific knowledge, can best exemplify general vocabulary but still the presentation must meet discourse requirements and shared expectations of the participants. Spoken interaction, on the other hand, depends very much on the participants, occasion, and context of a speaking situation. Depending on who the participants are, the language used will be more or less formal, with higher or lower degrees of politeness, as per the cultural norms and shared expectations on behalf of the participants. It will also be influenced by the occasion in which participants find themselves having a conversation and sharing the responsibility for constructing the meaning of the utterances. In the same vein, context will direct the exchanges according to the norms acquired by the participants, based on their previous experience and their both formal and informal education. The linguistic characteristic of a spoken interaction will vary as much as it will be influenced by the participants' experience and turn-taking skills; however, Green notes that during spoken interaction participants generally demonstrate simple grammar structures, often characterized by coordination (2014). Luoma goes further claiming that “the vocabulary of spoken interaction tends to be relatively generic and vague,” for example, “*the thing over there*”, rather than more precise words, such as “the blue bowl on the table”. (Luoma 2004 in Green 2014: 130).

Building upon the work of Luoma (2004), Tonkyn and Wilson (2004), and Hughes (2010), Green outlines a set of features characteristic of more or less proficient speech. These features are often found in descriptors used for rating spoken performance, i.e. in rating rubrics (Table 2.1).

Table 2.1: Features of more or less proficient speech (Green (2014: 131) based on Luoma (2004), Tonkyn and Wilson (2004), and Hughes (2010)

Less proficient speech	More proficient speech
Shorter and less complex speech units	Longer and more complex (e.g. more embedded) speech units
More errors per speech unit	Fewer errors per speech unit

Less and more limited use of cohesive markers (and, but, etc.)	More and more varied use of cohesive markers
Limited to use of common words	Use of more sophisticated and idiomatic vocabulary
Pauses linked to language search	Pauses linked to content search
Pauses within grammatical units	Pauses between grammatical units
More silent pause time	Less silent pause time
Shorter stretches of speech between noticeable pauses	Longer stretches of speech between noticeable pauses
Speed of delivery noticeably below typical native speaker rates	Speed of delivery not noticeably below typical native speaker rates

2.2.2.1.2 *Monitoring and revising*

Due to the nature of shared responsibility for the talk during the spoken interaction, participants feel obliged to help the meaning to be constructed and realized in accordance to their communication goals and language/social conventions. This process also involves monitoring the transmission of the message and revision, if *accuracy* is of primary concern. Sometimes it is sufficient for the addresser to see the face of their interlocutor to realize if there are any problems with the understanding of the intended meaning of the message or not. In case the remedy measures have to take place, their nature may vary concerning the problem detected during the transmission of the message. For example, the meaning can be affected at the phonological level, so that the remedy has to take place and the mispronounced units have to be corrected or repeated for better understanding of the message. Alternatively, the impediment may occur in relation to grammar, so that the remedy will tackle grammar issues. On the other hand, accuracy does not have to be the goal, so the participants in the interaction opt for fluency, and consequently they may disregard any inaccuracies emerging throughout the interaction, for the sake of fluency. In other words, the issues with accuracy do not have to “damage” the message enough for the interaction to take a break in order for the remedy measures to take the place (Green, 2014). Test developers should consider the above-mentioned issues when designing items to assess speaking tasks.

3 Testing

3.1 Testing communicative language ability

This chapter offers a brief overview of communicative language testing, including the history of its development. Next, the author discusses the nature of specific purpose (SP) language testing, which is considered to be communicative “by definition” (Douglas, 2000:19). This part of the discussion is relevant to understanding two concepts inherent to SP language testing: specific purpose target language situation and background (or topical) knowledge. Finally, the central role of tasks is discussed within the task-based approach to language testing, with relevance to demonstrating one’s language ability outside the educational setting.

3.1.1 Communicative language testing

Communicative testing developed under the influence of the model of communicative language ability, in the last two decades of 20th century, and has kept its place in the focus of testers’ attention ever since. Douglas argues that even in the 1980s, the topic of communicative language testing was not entirely new, because a decade earlier, language testers had been discussing “productive communication testing” (Upshur, 1971 in Douglas, 2000: 9). In 1990, Cyril Weir published his book *Communicative language testing*, in which he defines it as follows:

In testing communicative language ability we are evaluating samples of performance, in certain specific contexts of use, created under particular test constraints, for what they can tell us about a candidate’s communicative capacity or language ability. (Weir, 1990 in Douglas, 2000:9)

Based on this definition, it became apparent that language testers would have to base their test development decisions on several key terms: *communicative language ability*, *specific contexts of language use*, *test limitations*, and candidates’ *capacity*. Bachman defined communicative language ability in 1990, drawing language specialists’ attention to the use of language in particular contexts whose many features (for example, time and place, participants in communication, the topic, etc.) or characteristics inevitably affect communication that takes place in a given context. Consequently, the need to define the context of language use was motivated by practical considerations of determining those special characteristics of the context that need to be replicated in the corresponding testing situations. Bachman and Palmer insist that

a target language use situation and the test tasks sampled to represent this situation and its language tasks must have something in common in order to provide the link between the ability to respond to a test task and the ability to demonstrate the corresponding communicative behavior outside the testing situation (1996: 9). The issue with testing situations, however, refers to tests being artificial events, designed in order to elicit particular behavior. This is where the considerations of test constraints come into play, since the method used to elicit and assess a language performance inevitably affects that performance. The familiarity with test constraints is essential if testing is to claim overall construct validity in Bachman and Palmer's sense (ibid.). Finally, the last key term in Weir's definition, the one referred to as capacity, demonstrates what Widdowson described as "the ability to use knowledge of language as a resource for the creation of meaning" (1983 in Douglas, 2000:10). Douglas employs the meaning of capacity to explain language situations from the perspective of language users, considering their understanding of the context and language use in it as a key approach to assessing specific purpose language ability.

3.1.2 Specific purpose language testing

Douglas argues that there is no significant difference between communicative language tests and specific purpose language tests, and proposes that specific purpose language tests should be considered as a special case of communicative language tests (ibid). He defines a specific purpose language test as follows:

A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain.

Douglas (2000:19)

Douglas's definition of specific purpose language testing emphasizes the importance of specific purpose target language situation, the analysis of which will provide the basis for developing test tasks with corresponding characteristics. The concepts of situational and interactional authenticity are embedded in the definition, stressing the importance of authentic approach to testing specific purpose language ability. On the one hand, situational authenticity enables test developers to replicate the characteristics of specific purpose target language situation to the

testing context, whereas this type of authenticity alone is not enough if it is not complemented with authentic interaction between the characteristics of test takers and test task characteristics. The analysis of the context and language tasks is crucial to understanding specific purpose language ability because language performance varies with both context and test task resulting in the interpretations of a test taker's ability varying from situation to situation (ibid.).

The issue of test constraints is one of the attributes of communicative language tests, and since specific purpose language tests are communicative by definition, it is worth mentioning certain limitations inherent to specific purpose language testing. First, test results are supposed to generalize to some real life domain of target language use, demonstrating that test takers possess language skills to operate within this domain. In the case of specific purpose target language use domains, it is difficult to sample all tasks representative of the domain. Additionally, even in the case of narrowly defined contexts, test developers cannot easily sample and cover all variables that are inherent to specific purpose language situations. Second, in general purpose language tests, the issue of topical or background knowledge is referred to as a potential source of score contamination. In specific purpose language testing, background knowledge is a necessary component, because it can be argued that specific purpose language knowledge includes what Bachman and Palmer call *topical knowledge* – the knowledge of a field-specific subject matter, including specific purpose vocabulary (1996). If we argue that topical or background knowledge in Douglas's sense is a component of specific purpose language ability, this knowledge will be a part of the construct measured in a specific purpose language test (for a detailed discussion of background knowledge see 2.3.1 above).

3.1.3 Task-based (performance-based) language testing

In recent history of language assessment, test developers have been dealing with testing language ability in broad sense, which is known as construct-based approach to assessment, and testing language by focusing on language tasks and language use contexts within task-based approach to assessment. We will discuss constructs later in this thesis, but it is worth taking note of the fact that these two approaches are not conflicting but rather complementing, since in both cases test developers employ test tasks to assess the ability in question. However, there are two factors to determining whether an assessment follows a construct-based or a task-based approach to test design: (1) the role of tasks in the assessment, and (2) the purpose of assessment. These

two factors are interrelated in the sense that the purpose of assessment determines the relative role of tasks in test design. If the purpose of an assessment is to provide general information about a test taker's language ability, the test tasks will be designed in line with construct-based approach, resulting in test scores that place the test taker's ability at a certain level of proficiency. If test tasks are used to determine how well the test taker performs on a task in a certain context, their test scores will generalize to a specific target language situation outside the testing context. Bachman warns that this is a recommended course of action if characteristics of a target language use situation are easy to define (Bachman 2002, in Luoma, 2004:42).

There are two distinct advantages to task-based testing: directness of testing method and potentials to increased authenticity. Task-based approach in assessment emerged as a measure which will secure "more direct and more accurate testing because students are assessed as they perform actual or simulated real life task" (Brown and Abeywickrama, 2010: 16); hence, the alternative term – "performance-based" testing. As a positive outcome of task-based approach to testing, the assessment may claim to possess higher content validity, since performing on a task is a direct measure on the ability tested. For example, if a test intends to measure test takers' ability to participate in "small talk", test developers will design an interactive speaking tasks, requiring the participants to adhere to the social and linguistic norms inherent to what we know as "small talk". In responding to tasks pertaining to task-based approach, test takers are involved in an array of activities involving oral and written production, open-ended response type tasks, interactive task types, group task types (e.g. group presentation), etc. Task-based approach is not only task-centered, but also learner-centered in terms of the accountability for the assessment process, the freedom of choice, and the lack of strict structure, unlike in more traditional test task types. When they respond to a task which shares the characteristics of a non-testing situation task, test takers focus not only on their language abilities but also to the requirements of the target language use situation in terms of their specific role in it. Consequently, this may have a positive impact on situational and interactional authenticity of the task, contributing to the test's overall usefulness.

3.2 Types of tests based on the intended use of test results

In assessment contexts, test users are final users of scores derived at the end of a testing process; therefore, this term refers to different individuals, groups of individuals and institutions.

What they all have in common is that they require test scores/results in order to be able to make certain inferences and decisions regarding test takers. Speaking of individuals and groups of individuals, in educational contexts they normally refer to faculty, teachers, instructors, module co-teachers, and other individuals who want to know how successful individuals are in meeting learning objectives, or how successful syllabi are in achieving the goals set by an educational institution. When it comes to contexts other than that of education, individuals and groups of individuals who can take the role of test users refer to employers, managers, HR officers, employment consultants, etc. The other group of individuals who want information regarding test results can come from both educational and other contexts, meaning that they can be regarded as representatives of various institutions – enterprises, companies, state agencies, governmental statistical agencies, educational institutions, ministries, etc.

Test takers can also be regarded as test users, though they seldom need test results *per se*. For example, students, at all levels of education, are interested in knowing what their grades are in order to see how successful they are as students, how close they are towards meeting curricular requirements, and how far they are from graduating and earning a diploma or degree. In other contexts, outside academia or education, test takers are also interested in what they can do with test results, rather than in results themselves. For example, they might be taking tests for various purposes: immigration, employment, professional development, promotion. In other words, it is less important what the actual result is than what a person can achieve with it.

Test purpose is closely related to score interpretation, and the manner by which scores are interpreted depends on the way constructs in a particular assessment were defined. This relationship conditions the kind of tasks that are selected for a particular assessment, ensuring that the construct they cover matches course syllabus requirements, in case of educational domain, or the facets of target language use tasks, in the real life domain. The awareness of the relationship that exists between test purpose and intended use of test scores is crucial to understanding the correlation that exists between target language use domains and construct definitions in the contexts of two types of assessments that will be discussed below. Consequently, this correlation plays a role in determining whether an assessment is more or less authentic, since it is test purpose and intended use of scores that contribute to establishing authentic relationship between test tasks and target language use tasks.

3.2.1 Criterion-referenced tests (CRTs) vs. norm-referenced tests (NRTs)

There are two kinds of tests with regards to the intended interpretations of test scores: criterion-referenced tests and norm-referenced tests. The former draws upon curriculum/course syllabus which includes various learning objectives and outcomes, and each one of the outcomes is assessed in one manner or another in order to make sure that all test takers have mastered the same knowledge and skills. For example, in the context of language assessment, one of the learning outcomes may be ensuring that students have mastered the skill of recognizing the main idea of the recording in listening comprehension assessments. Mastery of a particular skill or a piece of knowledge is then taken as a criterion based on which inferences will be made regarding the student's progress. Accordingly, this approach to testing is known as *criterion-referenced testing*. Brown and Hudson define a criterion-referenced test (CRT) as “any test that is primarily designed to describe the performances of examinees in terms of the amount that they know of knowledge or set of objectives” (2002:5). Additionally, due to their formative nature, these tests are useful for any assessment situation within educational domains because each test taker can achieve a maximum score if they have mastered the full amount of knowledge as per the course syllabus. In other words, all test takers can get a score of 100 percent if they have mastered the course content entirely. Criterion-referenced tests are designed to provide feedback to test takers, and this feedback can take the form of grades, related to learning objectives, but it is often accompanied with a description of the performance mapping strengths and weaknesses in it (Brown and Abeywickrama, 2010:8).

The other kind of tests, based on the intended interpretations of test scores, rank students “along a mathematical continuum” proving their full potential in making selection decisions (Brown and Abeywickrama, 2010: 8). Test scores are interpreted in the form of a numerical score, or a percentile rank, showing a test taker's relative standing in comparison to others (ibid.). Test items in *norm-referenced testing* prove their distinctive value by differentiating between candidates so that both stronger and weaker candidates are easily identified. According to Brown and Hudson, NRTs are designed in such a manner that they include “items that about half of the students cannot answer correctly on average” (Brown and Hudson, 2002: 7). In other words, in criterion-referenced tests, test items are designed so as to show what test takers know, whereas norm-referenced test should point out what it is that weaker candidates do not know

(Brown and Hudson, 2002: 7). The purpose of administering norm-referenced tests is often to make important decisions regarding successful candidates (employment, promotion, award, etc.); therefore, NRTs are applied in proficiency testing and testing for selection purposes. Since they do not match any particular course syllabus, and they do not foster improvement or further learning, NRTs are summative and discriminative by nature. Such are proficiency tests, for example TOEFL or IELTS; they do not cover any particular learning objectives and are not related to any specific course material. Their discriminative nature helps test users make decisions regarding candidates who take the tests. For example, high-scoring candidates are admitted to the course or granted a scholarship; whereas low-scoring candidates may be advised to enroll in a foundation program and improve their English language skills before they can be admitted to an undergraduate program. Although both criterion-referenced and norm-referenced tests can be administered in educational domains, only norm-referenced tests can be successfully applied in the real life domain, such as the one related to industry or economics. If the purpose of an assessment is to identify and hire the highest-scorers and top candidates for a position, it is norm-referenced testing that informs such decisions.

3.3 Test usefulness: qualities of a language test

The most important consideration in test design and development is whether the test will be useful or not; in other words, test developers and test users need to know whether the test is useful for its intended purpose or not. This consideration emphasizes the importance of test qualities, which at the same time determine and define its usefulness. Although it may go without saying that a language test should be useful, this usefulness has to be demonstrated and proved in a certain way. To ensure test usefulness, Bachman and Palmer propose the following model:

$$\text{Usefulness} = \text{Reliability} + \text{Construct validity} + \text{Authenticity} + \text{Interactiveness} + \\ \text{Impact} + \text{Practicality}$$

According to this model, test usefulness is seen as a function of different qualities “all of which contribute in unique but interrelated ways to the overall usefulness of a given test” (1996: 18). Although all qualities of test usefulness are to find their place in the overall evaluation of test usefulness, there are several issues that may make this evaluation difficult. First of all, these

qualities are demonstrated in different ways, and although they are complementary, in certain assessments some qualities will be more prominent than the others. Second, test qualities cannot be evaluated independently from one another, because each one of them will be represented to a certain extent in every assessment. It is impossible to “prescribe” an ideal and general balance that will apply to all testing situations because test purposes are different, and each test will be evaluated by its own merits. Third, these qualities are interrelated so test developers should be watchful from the beginning of the design process in order not to ignore any one quality, or to maximize any one at the expense of the others.

The authors of the model state that there are three guiding principles that help operationalize the model of test usefulness in any particular language test (ibid.):

Principle 1: It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.

Principle 2: The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.

Principle 3: Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation.

For a language test to be useful, it should be developed considering its intended purpose, test takers, and a specific situation in which the test takers will be using the language or its components assessed in the test.

Building on the work of Bachman and Palmer, other authors elaborate on the qualities of a language test in various contexts. Brown and Abeywickrama, for example, state that there are five “cardinal criteria for testing a test”, especially in the context of classroom assessment – practicality, validity, reliability, authenticity, and washback (2010:25). The purpose of these criteria is to help test developers find answers to the following questions (ibid.): “Can it [the test] be given within appropriate administrative constraints? Is it dependable? Does it accurately measure what [test developers] want to measure? Is the language in the test representative of real life language use? Does it provide information that is useful to the learner?” Discussing a plan for evaluating the qualities of good testing practice in the context of assessing languages for special purpose (this plan is a part of a test specifications document, see Chapter 5.3.1 below), Douglas refers to the following principles for ensuring the test usefulness (2000:112):

- 1) validity – the interpretations that can be based on test performance and test scores;
- 2) reliability – the consistency and accuracy of the process of measuring skills in a test;
- 3) situational authenticity – the relationship between language tasks in the target language situation and the test tasks;
- 4) interactional authenticity - the extent to which test takers' communicative ability is engaged by test tasks;
- 5) impact – the influence that test has on test takers, instructors, and educational systems; and,
- 6) practicality – the potential limitations caused by practical considerations, such as budget, administrators, available personnel, and institutional policies.

What all the authors above have in common is the view that the qualities of test usefulness are the guiding principles determining how useful a test will be, including the inferences made based on its results. Given the importance of decisions made according to the test results in the context of English for specific purpose assessments, the following chapters discuss respective test qualities.

3.3.1 Reliability

The scores obtained on a reliable test are consistent across test administrations. For example, if the same test takers demonstrate a poor performance on one occasion, they are expected to demonstrate the same or similar performance on a different occasion, provided they are given the same test, or the parallel form of the same test. Essentially, reliable tests are dependable in that they provide the correct information regarding test takers' language ability. In the context of classroom assessment, Brown and Abeywickrama identify the following factors that can affect the quality of reliability: *student reliability*, *rater reliability*, *test administration reliability*, and *test reliability* (2010: 29). The issue of *student reliability* in any particular testing situation refers to possible fluctuations in student's performance on the test. These fluctuations usually result from the student's physical and mental state at the moment of taking the test. For example, if students take a test when they are not feeling well, a poor performance may be attributed to their physical condition preventing them from giving their maximum in responding to test tasks. *Rater reliability*, on the other hand, is related to differences that may exist between ratings of the same performance assigned by different assessors, or, in

some cases, they can refer to differences in ratings provided by the same assessor. Accordingly, we may distinguish between *inter-rater reliability* and *intra-rater reliability*. The former refers to a testing situation where the same performance is judged by more than one person. In reliable assessments, the same performance is supposed to yield the same or about the same scores. This is particularly significant in the context of subjectively scored assessments, such as the case of assessing written or spoken performance. *Intra-rater reliability* refers to internal factors related to the same assessor on occasions when they rate the same quality of performance. Ideally, the same performance should yield the same scores, but this is not always the case. There are various strategies that an institution can take in minimizing the effect of raters' subjectivity in rating test takers' performance. For example, rater trainings and standardization sessions are organized so that raters' performance is consistent and dependable. *Test administration reliability* refers to physical conditions on the occasion of a test administration. Brown and Abeywickrama identify several sources of unreliability with this regard: lights, noise, poor state of photocopied testing material, temperature, etc. (ibid.). Finally, test reliability refers to the test itself, its tasks and instructions, organization and time allocated to responding to test tasks. In classroom assessments, poorly written items are particularly problematic. Brown and Abeywickrama suggest that problems most often occur in the case of subjective tests with open-ended question types. This problem can be mitigated by the use of well-developed rating rubrics and sample performances helping assessors identify strengths and weakness in test takers' performance (ibid.). Discussing the quality of reliability, Bachman and Palmer consider reliability as a function of the consistency of scores from "one set of tests and test tasks to another" (1996: 20). They primarily think of tests as sets of task characteristics, so that reliability is observed as a consistency between two administrations of the same test, applied to the same group of test takers. In the norm-referenced assessment situation, for example, the rankings are supposed the level of mastery of a desired knowledge or a language skill. If the same test takers take the same norm-referenced test, their respective standing on the rank list should be more or less the same. If this is not the case, however, then it is quite possible that there is a problem in the test characteristics not yielding in the parallel test forms. However, one should not exclude other source of unreliability (for example, issues regarding test administration, or problems related to test takers, such as health issues, etc.). Bachman and Palmer argue that "reliability is essentially

the quality of test scores”, because the inferences based on test scores depend on their consistency, i.e. reliability.

3.3.2 Validity

Traditionally, validity has been considered the most important criterion of a test usefulness, since it refers to its ability to measure what it is intended to measure (Messick, 1989). At a first glance, this claim seems redundant, since testers develop tests in order to measure a specific skill or a component of language ability; however, it can be argued that in some cases tests lack validity, meaning that inferences based on test results are not quite meaningful and appropriate interpretations of test takers’ language ability. As is the case with many other issues inherent to language assessment, there is no general consensus as to how a broad understanding of validity can be subdivided into its constituent parts. Bachman and Palmer, for example, argue that construct validity is one of the six qualities of test usefulness, defining it as the extent to which interpretations of test scores are meaningful and appropriate (1996: 21). Brown, on the other hand divides validity into three broad categories: content, criterion-related, and construct-validity (1996: 231-249), whereas Brown and Abeywickrama further develop the notion of validity, stating that it includes several sources of evidence (content-related evidence, criterion-related evidence, construct-related evidence) and adding to it two more types of validity (consequential and face-validity). It should be noted, however, that some other authors, such as Bachman and Palmer (1996) and Douglas (2000) consider the issue of consequential validity as a separate test quality which they call impact, referring to the influence a test may have on learners, teachers, institutions, and educational systems.

Brown and Abeywickrama emphasize the importance of validity in all testing situations, because they argue that a valid test (2010: 30):

- measures exactly what it proposes to measure
- does not measure irrelevant or “contaminating” variables
- relies as much as possible on empirical evidence (performance)
- involves performance that samples the test’s criterion (objective)
- offers useful, meaningful information about a test taker’s ability
- is supported by a theoretical rationale or argument

Ideally, a valid test contains all the attributes cited above, however, according to McNamara and Roever (2006) it is impossible to establish absolute measure of validity. Validity is an ideal,

something that should be achieved in a test, but nobody can state that the inferences based on their test results are valid, unless they support the claim by evidence. Additionally, Messick underlines that it is important to bear in mind that “validity is a matter of degree, not all or none” (1989:33).

The evidence, supporting the claim that test results are valid, may come from the following sources: content, criterion, and construct. We may mention as well that there is another type of evidence known as *face validity*, although it should be noted that this type of evidence is what Bachman and Palmer refer to as a “superficial factor”, dependent only on the eye of the perceiver (1990: 285-289). Face validity is established often by observation, on behalf of test takers, and their perception of the test as an instrument that is intended to measure their language ability. Despite the fact that some authors consider this type of evidence superficial and useless, it can be argued that test takers’ performance can be influenced by their perception of the test. For example, if they face a new item format on the test itself, their performance may be affected by the fear of the unknown test item (at the same time, this may cause student-related unreliability, which often affects the validity of scores on a test). To mitigate the effect of unreliability and to increase students’ perception of fair testing process, Brown and Abeywickrama propose that test developers should (2010: 35):

- use well-constructed, expected question formats with familiar tasks,
- create tasks that can be responded to within allotted time limit,
- select test items that are clear and uncomplicated,
- write directions that are easy to follow,
- choose the tasks that students are familiar with,
- create tasks which reflect the coursework (content validity),
- use task whose difficulty levels are reasonably challenging and balanced.

It should be noted, though, that test validation process is not restricted to any particular stage of test development; rather, it is an on-going and iterative process that can be applied to any stage of the process. Some authors do not even make the difference between validity and construct validity, so the quality of validity is often referred to as the construct validity (like for example in the work of Bachman and Palmer, 1996). However, the author of the thesis believes that the distinction among various sources of evidence of validity may help readers with understanding the complexity of the concept of validity. Below, we will discuss several types of evidence that can be provided in support to the process of test and score validation: content-related evidence, criterion-related evidence, and construct-related evidence.

3.3.2.1 Content-related validity

Collecting evidence in order to prove that a test has content validity includes a number of strategies that focus on the test's content, ensuring that it covers course objectives, for example. In classroom assessments, a test is an instrument sampling the subject matter outlined by a course syllabus. The test itself serves as an instrument which will measure to what extent test takers have mastered the points from the syllabus, but to do so, the test items should be closely related to those points. This brings up the difference between **direct** and **indirect testing**. In the case of the former, a particular skill is tested by asking test takers to actually demonstrate it in their performance. For example, if test developers' intention is to assess if test takers know how to use a "hook" at the beginning of an oral presentation, they will instruct them to start with any kind of input that can draw the audience's attention, and then it is upon the assessors to decide how meaningful and appropriate the hook was. Brown and Abeywickrama propose that test developers can identify content-validity observationally, provided the achievement being measured is well defined based on the course objectives. Additionally, they suggest adhering to direct testing as much as possible, as this is a sure proof way to assess desired knowledge or a skill and map the observed behavior on the list of learning outcomes (2010:30-32).

3.3.2.2 Criterion-related validity

Criterion-related validity refers to any evidence that will link the classroom test to another, external, well-respected measure of the same ability. In other words, a classroom test is a set of samples related to a certain criterion, for example a point of grammar in communicative use. Such test will prove to have a criterion-related validity if its results are compared to some other measure of the same criterion. In case of large-scale high-stake assessments, if an organization decides to create such proficiency test, it will administer its own test and then some well-known proficiency tests, such as TOEFL, to the same group of test takers and then corroborate that both tests are proficiency tests measuring the same objectives by comparing the results. If there is a high degree of correlation between the two sets of results, achieved by the same test takers, the organization's proficiency test may be claimed to possess criterion validity. The latter is, at the same time, an example of the test's *concurrent validity*, which refers to test takers' possessing the same "amount of knowledge" since both tests are administered at about

the same time. Another example of concurrent validity is when test takers achieve relatively high results on the final exam at the end of Semester 1, and their actual language proficiency at the beginning of Semester 2 corroborates the results on the exam. There is another quality of criterion-related validity that is worth mentioning here. It is related to its quality to predict future performance, which is important in placement and achievement tests, so this kind of validity is known as *predictive validity*. As the term itself suggests, the value of this kind of validity evidence lies in predicting whether test takers are likely to achieve success in the future.

3.3.2.3 Construct-related validity

Construct validity is often regarded as a central point in collecting evidence to prove the validity of test scores and inferences based on them. As a quality of test usefulness, construct validity aims at proving the following:

- 1) test scores reflect the language ability that the test intends to measure, and
- 2) test scores are evidence of the test taker's language ability to perform the TLU tasks.

The purpose of a language test is to measure test takers' language knowledge or skills, demonstrating that test scores reflect the degree of their possession of knowledge or their mastery of a particular skill. If the test fails to do so, it is no longer an indicator of the test takers' language ability, despite the test developers' intentions or efforts. Brown, for example, identifies thirty-six threats to test reliability, warning that these directly influence the validity of its results. He divides these threats into five categories: environment of the test administration, procedural failures related to test administration, test takers, scoring method, and the quality of test items/test as a whole (1996: 188-192). Knowing that reliability can affect the test validity, test developers need to focus on the issue of measuring the desired language ability, in order to provide evidence that test scores can be used as "an indicator of the abilities, or constructs" they want to measure (Bachman and Palmer, 1996: 21). To do this, their starting point will be to define the construct their test intends to measure so that this definition can be used as the "basis for a given test or test task and for interpreting scores derived from this task" (ibid.).

Ideally, test scores prove their value outside the testing context, i.e. in the target language use domain, or to what Bachman and Palmer refer to as the "domain of generalization" (ibid.). In other words, the domain of generalization refers to the target language use domain and the tasks

to which test tasks correspond. Test takers are assessed so that they can demonstrate their language ability, and the scores that they receive upon the test administration demonstrate how successful they are likely to be at performing on the corresponding language tasks in the target language use domain.

Bachman and Palmer argue that in determining the construct validity of any given score interpretation, test developers and test validators need to consider both the construct definition and test task characteristics. The former is important because it defines the ability that will be observed and measured by the test, while the latter is relevant to the target language use domain, since it shows the extent to which test tasks correspond to the tasks in the target language use domain. This is also known as the quality of authenticity in Bachman and Palmer (1996), or one aspect of authenticity, called situational authenticity in Douglas (2000). In addition to situational authenticity, it is important to analyze test task characteristics in order to determine the degree to which they may engage the test takers' language ability or its components. Bachman and Palmer call this quality "test interactiveness" (1996: 22), considering it one of the six qualities in the process of determining test usefulness. Douglas, on the other hand, argues that authenticity is manifested as situational (showing how test tasks and the target language use tasks share the same set of characteristics), and interactional (demonstrating how the authentic test task characteristics engage the appropriate discourse domain in the test takers). Once the appropriate discourse domain has been engaged, it helps test takers interact with the task in the same way language users respond to a language task outside the testing context, i.e. in the target language use situation (2000:112).

In Bachman and Palmer's sense, construct validation is an ongoing process, and the types of evidence discussed above are pieces of the mosaic that are collected and put together in support to the claim that interpretations of particular test scores are valid.

3.3.3 Practicality

Practicality is a test quality that deals with constraints imposed on the process of test development, its administration, and the scoring method applied so to produce test scores. This issue, as the name suggests, refers to practical considerations that any institution, language unit, or a teacher should bear in mind, if they are to meet the requirements of cost- and time-effective

test administration. Brown and Abeywickrama suggest that test developers should consider the following attributes of practicality (2020: 26):

A practical test:

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

These considerations are easy to apply to any testing context, since the issue of practicality applies to any test administration. For example, if two oral assessors are hired to assess the spoken performance of 200 students on the same day, it is easy to deduce that due to fatigue, inter- and intra-rater reliability issues that may occur in given circumstances test administration, and the validity of test scores can be called into question.

Bachman and Palmer emphasize the cyclical nature of test development process, stating that the issue of practicality can be applied to each one of them. In simple terms, the authors define practicality as “the relationship between the resources that will be required in the design, development and use of the test and the resources that will be available for these activities” (1996: 36). If test design, development and its administration require more than the available resources, the test is likely to be impractical. Speaking of resources, Bachman and Palmer classify resources into three general types: human resources, material resources, and time. *Human resources* refer to test developers, item writers, test administrators, invigilators, support staff, etc. *Material resources* encompass space, equipment and materials required to complete the testing process. *Time* includes the time required to complete the development process, from the beginning of the process to the moment when scores are reported, as well as the time allocated for specific tasks (p.37).

3.3.4 Authenticity

One of the most distinctive characteristics of communicative language tests is their claim to feature authentic test tasks. However, this is a rather bold claim, because authenticity is not a quality that is easily measured and proved, especially since language tests often contain contrived language and manipulated stimuli aimed at eliciting certain responses from test takers. In addition, it can be argued that authenticity is a matter of degree, the higher the degree the

more authentic the assessment is. Bachman and Palmer define authenticity as “the degree of correspondence of the characteristics of a given language test task to the features of a target language task” (1996:23). Douglas recognizes this correspondence between the two set of tasks as situational authenticity, and in addition to it, he proposes that authenticity is also the quality that resides in test takers. Their interaction with the characteristics of test tasks determines the extent to which they are involved in solving them, hence the term interactional authenticity (2000:112). Authenticity as a key concept of interest for this thesis will be thoroughly discussed in Chapter 3.

3.3.5 Impact

Test impact is a quality of language tests linked to the values, judgments, and consequence their administration and use have on individuals as well as the society as a whole. Bachman argues that “tests are not developed and used in a value-free psychometric test tube” because they are intended to serve other purposes – those imposed by a curriculum, educational system and society (Bachman, 1990 in Bachman and Palmer, 1996: 30). Consequently, there are certain values and goals inherent to the process of administering and taking the test. Bachman and Palmer point out that test impact operates at two levels: micro and macro. At a micro level, individuals are affected by testing practices, whereas at a macro level, tests can have consequences for educational systems and society (ibid.). An important aspect of impact is *washback*, which Hughes (1989) defines as “the effect of testing on teaching and learning,” which can be either beneficial or harmful (in Bachman and Palmer, 1996: 30) and may occur both at micro- and macro levels. However, it is worth noting that, unlike washback that occurs at both micro- and macro-levels, *backwash* affects only individuals, in either positive or a negative way (for more on backwash see, Green 2003, and Weir, 2005). Test developers and administrators should be aware of the possibility that their tests may put test takers at disadvantage in order to minimize its potentially negative effects. Brown and Abeywickrama talk about test impact in terms of consequential validity that includes three important sets of considerations: “accuracy in measuring intended criteria, its [a test’s] effect on preparation of test takers, and the (intended and unintended) social consequences of a test’s interpretation and use” (2010: 34). The issue of accuracy is as relevant to the context of language assessment as to any other context which deals with measurement instruments. If test scores are to be used for their

intended purpose, they should be arrived at by using as precise “measurement” techniques as possible. This quality is relevant to all assessments, but it seems to be especially significant in high-stakes assessments (for example, in entrance or final exams, in internationally recognized standardized tests of proficiency, as well as in tests that have gate-keeping purposes, as is the case with the tests administered to suit the purpose of immigration), given the gravity of the decisions based on their results. Speaking of a test’s impact in terms of preparation of test takers, this issue is closely related to the consideration of test fairness. In other words, this consideration deals with test takers’ familiarity with the context presented by test tasks, test task types and formats, test takers’ accessibility to coaching and preparation courses, etc. Finally, if we accept Bachman’s claim that tests can affect individuals and the whole society both directly and indirectly, the issue of social consequences caused by the test should be taken into consideration (for more on ethical considerations related to test fairness, see Milanović and Milanović, 2013).

3.4 Assessing speaking skills

This chapter offers an overview of speaking skills, target language use tasks, and speaking test tasks corresponding to the real life tasks. First, the author outlines the differences between micro- and macroskills of speaking. Second, the author discusses the meaning of a task and context outside the testing context, paving the way for better understanding of speaking tasks. Finally, the discussion ends with a brief outline of the most common speaking task types.

3.4.1 Speaking skills

Speaking assessment is aimed at eliciting test takers’ speaking skills in the target language via an appropriate test method developed in accordance with the purpose of an assessment and the intended use of test results. Although no one can deny that test method is crucial to eliciting the knowledge and skills that are to be tested, Alderson et al. warn that the very method testers use to test a language ability may “affect the student’s score”; and this is known as “the method effect” (1995:44). In the essence of the “test method effect” lies a possibility that test takers develop the skill of solving particular task types (for example, some test takers study solving multiple-choice tasks, becoming skillful in distinguishing between distracters and the correct answer), leaving test users in ignorance whether their scores really represent their language skills and knowledge or their ability to solve the tasks in question. This

is also known as test wiseness (ibid.). To minimize the test method effect, various authors advise test developers to clearly identify and define the object of measurement – the skills and knowledge that are relevant to a certain testing purpose. In the case of speaking assessment, the starting point is to identify the relevant skills that are to be observed and tested. Brown and Abeywickrama distinguish between **micro-** and **macroskills** of speaking. The microskills of speaking refer to producing “the smaller chunks of language such as phonemes, morphemes, words, collocations, and phrasal units”, whereas the macroskills refer to producing spoken units by combining larger elements such as “fluency, discourse, function, style, cohesion, nonverbal communication, and strategic option” (2010:142). They provide a list of micro- and macroskills that test developers can refer to in the process of test developing and item writing. Language learners start by developing microskills, and as they progress they devise the skill of saying the same thing in different ways. Combining their microskills they continue their progress towards proficiency by engaging into increasingly complex units of oral production, when they start threading on the ground of developing and using macroskills of speaking (see Table 3.1 below).

Table 3.1: Micro- and macroskills of oral production (Adapted from Brown and Abeywickrama, 2010: 142-143)

<p>Microskills</p> <ol style="list-style-type: none"> 1. Produce differences among English phonemes and allophonic variants. 2. Produce chunks of language of different lengths. 3. Produce English stress patterns, words in stressed and unstressed positions, rhythmic structure, and intonation contours. 4. Produce reduced forms of words and phrases. 5. Use an adequate number of lexical units (words) to accomplish pragmatic purposes. 6. Produce fluent speech at different rates of delivery. 7. Monitor one’s own oral production and use various strategic devices – pauses, fillers, self-corrections, backtracking – to enhance the clarity of the message. 8. Use grammatical words classes (nouns, verbs, etc.), systems (e.g. tense, agreement, pluralization), word order, patterns, rules, and elliptical forms. 9. Produce speech in natural constituents: in appropriate phrases, pause groups, breath groups, and sentence constituents. 10. Express a particular meaning in different grammatical forms. 11. Use cohesive devices in spoken discourse. <p>Macroskills</p> <ol style="list-style-type: none"> 12. Appropriately accomplish communicative functions according to situations, participants, and goals.

13. Use appropriate styles, registers, implicature, redundancies, pragmatic conventions, conversation rules, floor-keeping and –yielding, interrupting, and other sociolinguistic features in face-to-face conversations.
14. Convey links and connections between events and communicate such relations as focal and peripheral ideas, event and feelings, new information and given information, generalization and exemplification.
15. Convey facial features, kinesics, body language, and other non-verbal cues along with verbal language.
16. Develop and use a battery of speaking strategies, such as emphasizing key words, rephrasing, providing a context for interpreting the meaning of words, appealing for help, and accurately assessing how well your interlocutor understands you.

Brown and Abeywickrama propose that test developers refer to the sets of skills listed above, when they set about designing test tasks to assess speaking skills.

3.4.2 Speaking tasks

John, B. Carrol, in his book *Human Cognitive Abilities*, defined a task as “any activity in which a person engages, given an appropriate setting, in order to achieve a specifiable set of objectives” (1993: 8). Additionally, Carrol underlines the following aspects relevant to language use tasks:

- the individual must understand what sort of result is to be achieved, and
- the individual needs to have some idea of the criteria by which performance will be assessed. (Carroll, 1992 in Bachman and Palmer, 1996:44)

Building on Carrol’s work, Bachman and Palmer propose that a language use task is “an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation” (ibid.). Defining communicative tasks, Nunan states that they involve “input, goals, roles, and setting” (1993 in Luoma, 2004:31). Accordingly, speaking tasks can be defined as tasks responding to which individuals use language to achieve particular goals in a specific speaking situation (Luoma, 2004.). It is evident that all authors agree that language tasks in general take place within a *context* that guides the speakers in their attempt to achieve their particular communicative goals. To understand how particular goals are achieved in a particular context, it is essential to understand the context itself.

Context is one of the important concepts in language use, and consequently in language learning and testing. Broadly defined, context covers “the linguistic, physical, psychological, and social dimensions” in a language use situation (Luoma, 2004:30). Luoma argues, that apart from the talk itself, context covers all other aspects of a speaking situation, such as, the place, time, roles of interlocutors, their language experiences and particular communicative goals, etc. (ibid.). Later in this work, we will talk about target language use situation whose certain characteristics overlap those of context. However, it is important to bear in mind that even when they identify and closely describe a certain target language use situation, test developers can seldom predict how dynamic aspects of the context (speakers’ knowledge, attitudes, expectations, language use) will evolve to a detail. On the other hand, some other, more static aspects and their characteristics can be engineered tasks whose outcomes can be predicted.

The starting point in speaking test task design is to decide what knowledge and abilities test takers are supposed to demonstrate. Once the construct of speaking has been identified, test developers refer to target language use situations to delineate oral performance that corresponds to speaking assessment goals. An important step pertaining to this process is identifying the type(s) of talk that speakers demonstrate in speaking situations. If test tasks are supposed to reflect the real life tasks, test developers have to consider what it is that speakers do with spoken language outside a testing context.

In the testing literature, there are two focal points when it comes to looking at what speakers do with language tasks: conveying information and performing an action. According to Brown and Yule (1983), speakers organize information in different ways in order to deliver it in what they call informational talk. In particular, they identify four types of informational talk organization: *description*, *instruction*, *storytelling*, and *opinion-expressing/justification*. The authors’ intention was to categorize between various types of talk, starting with least difficult and ending with the most difficult talk type (in Luoma, 2004:31). The intention behind such division was to help test developers select the task according to the intended level of difficulty; however, from today’s perspective, description, for example, can be quite complex and intricate, involving well-developed vocabulary and grammar structures, and involving higher-order skills, so the appropriate level of proficiency should be taken into consideration if task types are to follow Brown and Yule’s classification. In 1987, in his book *Speaking*, Bygate divided

information-related talk into two broad categories: *factually-oriented talk* and *evaluative talk*. Factually-oriented talk includes speaking activities such as: description, narration, instruction, and comparison. Evaluative talk refers to the following: explanation, justification, prediction and making a decision. Test tasks, following the above categories, can be developed to focus on one or more subcategories, or to involve a combination of several of them; however, Bygate warns that test takers' performance on different task types may vary significantly (in Luoma, 2004). In other words, a test taker may be skilled in delivering performance that requires narration, but that does not mean that they are equally good at tasks requiring justification, for example.

Test developers can refer to the type of information speakers convey by language tasks, as in the paragraph above, or they can observe the tasks from the perspective of actions the speakers perform when they use language. This approach was influenced by Austin's research in *speech acts* (1968), which confirmed that people use language not only to convey a message, but also to perform a certain action, such as place an order or confirm something (in Luoma, 2004: 33). The speech acts theory influenced van Ek's (1975) and Wilkins's (1976) work on functional syllabus, which shifted the traditional focus on grammar to the language functions learners need to develop in order to perform tasks in real life, thus paving the way to communicative syllabus in the 1980s. The results of the shift in focus are best seen in the Common European Framework of Reference, which combines the aforementioned approaches (Council of Europe, 2001).

The authors of the Common European Framework of Reference (CEFR), arguing the use of spoken (and written) discourse in communication for particular functional purpose, state that spoken discourse is mainly interactional. From the beginning until the end, one initiative causes a response, followed by further conversational exchanges until participants have fulfilled their communicative goals. By engaging in communication, the participants use certain structures, combining them in the order that follows formal and/or informal "patterns of social interaction "schemata)"(p. 125). According to Council of Europe, functional competence can be divided into two categories: microfunctions and macrofunctions. Microfunctions refer to limited spoken production – short utterances, normally taking place during the course of conversation. The table XXX below lists some of the examples of microfunctions:

Table 3.2: Functional competence – microfunctions (Adapted from Council of Europe, 2001: 126)

<ul style="list-style-type: none">- imparting and seeking factual information:<ul style="list-style-type: none">• identifying• reporting• correcting• asking• answering - expressing and finding out attitudes:<ul style="list-style-type: none">• factual (agreement/disagreement)• knowledge (knowledge/ignorance, remembering, forgetting, probability, certainty)• modality (obligations, necessity, ability, permission)• volition (wants, desires, intentions, preference)• emotions (pleasure/displeasure, likes/dislikes, satisfaction, interest, surprise, hope, disappointment, fear, worry, gratitude)• moral (apologies, approval, regret, sympathy) - suasion:<ul style="list-style-type: none">• suggestions, requests, warnings, advice, encouragement, asking help, invitations, offers - socializing:<ul style="list-style-type: none">• attracting attention, addressing, greetings, introductions, toasting, leave-taking, etc.
--

Macrofunctions are related to the categories of spoken discourse that consist of shorter or longer sequences of sentences. They share the same set of functions that Bygate identified in 1987, and, at the same time involve an extended production in Bachman's sense (1990). The following are examples of macrofunctions (COE, 2001: 126):

- description
- narration
- commentary
- exposition
- exegesis
- explanation
- demonstration
- instruction
- argumentation
- persuasion, etc.

Both kinds of functions are employed in communication by the means of schemata, or patterns of social interaction that underlie communication. The participants in communication are aware that, for example, a question requires an answer, or that a response follows a greeting, etc. (ibid.).

The considerations of language functions, context of communication, and the purpose of assessment are what test developers start with in order to create test tasks which assess test takers' ability to use the target language. It should be noted though, that productive skills also reflect the concurrent ability to use receptive skills (such as reading and/or listening), since the stimulus normally comes through receptive channels.

3.4.2.1 Speaking test tasks

Speaking tasks are test developers' tools used to operationalize the intended construct in a particular speaking assessment. There are several important considerations that test developers need to bear in mind in order to make important decision regarding test design and item writing: (1) how to assess test takers – one at a time, in pairs, or in groups; (2) pedagogic tasks or real life tasks; and (3) construct-based or task-based assessment of speaking.

First, in speaking assessments, test takers are most commonly examined individually, but depending on a situation, they can be tested in pairs, or even in groups. Consequently, the corresponding tasks are designed as *individual, pair, and group tasks*. Each task type has its advantages and disadvantages related to the qualities of test usefulness. For example, in terms of practicality, group tasks require the least time to administer, however they may involve issues related to rater reliability in assessing group performance. Second, in speaking assessment task design, test developers face a choice between “pedagogic” tasks, or what Luoma refers to as “language-focused” tasks (2004:40), and real life tasks (which Losada et al. refer to as *authentic tasks*, 2017), corresponding to language tasks in target language domains. The purpose of pedagogic tasks is to reinforce learning of language structures and functions, whereas real life tasks prove their value by corresponding to non-testing contexts. Nunan claims that real life tasks “require learners to approximate, in class, the sorts of behavior required of them in the world beyond the classroom” (2001:40). Some authors refer to real life tasks as to *authentic tasks* claiming that their value lies in preparing learners for tasks in the real life by helping them

“replicate or rehearse the communicative behaviors which will be required of them” outside the language classroom (McGrath, in Losada et al., 2017:92). McNamara distinguishes between *strong performance testing* and *weak performance testing*. In the case of the former, real life tasks, ensuing from a careful target language situation task analysis, replicate the target language situation task, and is judged by the real life criteria. In the case of weak performance testing, “having enough language ability” is sufficient for scoring well in a test (McNamara, 1996 in Luoma, 2004:40-41). Third, test developers need to decide whether their approach to task development will be construct-based or task based. In the case of the former, they will define the ability they want to measure and ensure its reliable and valid measurement. In the case of the task-based approach to task development, test developers have to identify and closely examine target language situation task, ensuring that its characteristics are reflected in test tasks. The latter approach is deemed very convenient in professional contexts (Douglas, 2000).

3.4.3 Speaking test task types

The discussion that follows offers a brief overview of the most common speaking tasks employed in learning settings. Tasks discussed below are classified according to their structure and the type of performance that is expected of test takers.

3.4.3.1 Structure: structured, open-ended, role-play

Luoma distinguishes between *structured* and *open-ended* speaking tasks in terms of “the relative amount of **structure** that the tasks provide for the test discourse” (2004: 47-48). In the case of the former, test items are designed so as to elicit narrow aspects of spoken production in controlled conditions, with limited expected response options. Such highly structured speaking tasks perform the same function as multiple-choice items in pen-and paper assessments. Open-ended tasks, on the other hand, give test takers the possibility to answer in a number of different ways. First of all, open-ended tasks can be *short* or *long*, depending on the purpose of assessment. Next, they can be classified according to the *discourse* type that they refer to; for

example, they may involve some of the following: description, narration, instruction, comparison, explanation, justification, prediction, and decision (ibid.). Some of these discourse types are employed even in high-stakes tests of speaking. For example, in the Internet-based Test of English as a Foreign Language, there are two independent speaking tasks that involve open-ended responses based on the discourse types mentioned above. In the first independent task, test takers are expected to demonstrate the ability to talk about a personal preference when they are given a choice to talk about certain categories – for example, people they find important, events and activities, etc. They are expected to demonstrate the ability to describe, explain, justify, etc. (see Example 3.1 below).

Example 3.1: Personal preference

Describe an ideal marriage partner. What qualities do you think are most important for a husband or wife? Use specific reasons and details to explain your choices. (Barron's, 2006)

In the second task (see Example 3.2 below), the idea of a personal choice is further developed, because test takers are required to make a choice and defend it while choosing between two contrasting courses of action or behaviors. Accordingly, they are expected to demonstrate the ability to describe, explain, justify, compare, contrast, decide, etc.

Example 3.2: Making a choice

Some people like to watch the news on television. Other people prefer to read the news in a newspaper. Still, others use their computer to get the news. How do you prefer to be informed about the news and why? Use specific reasons and examples to support your choice. (Barron's, 2006)

If the intention of open-ended task developers is to simulate real life tasks, in that case test tasks engage test takers in a *role-play*. According to Luoma, these tasks simulate the characteristics of a real life context (or target language use situation in Bachman and Palmer's sense, 1996) in order to provide inferences on test takers' ability to perform language tasks required in that context. She identifies three kinds of contexts in which role-play task can be applied: professional, social, and the context of providing a service (e.g. going to a restaurant). In response to role-play tasks, test takers assume certain roles and deliver spoken performance in

accordance with conventions of a given situation. It is interesting to note that presentations combine characteristics of role-play and discourse type tasks since they “combine the elements” of social conventions and the conventions of discourse that is involved in the execution of the task (Luoma, 2004: 49).

3.4.3.2 Type of performance: imitative, intensive, responsive, interactive and extensive

Brown and Abeywickrama, on the other hand, outline five main types of oral assessment tasks in relation to the **type of performance** test takers are expected to deliver, dividing tasks as follows: *imitative*, *intensive*, *responsive*, *interactive* and *extensive*. *Imitative* speaking tasks are concerned with phonetic level of oral production. In the light of communicative assessment, they are argued not to have communicative value. However, Brown and Abeywickrama point out that research has shown that “an overemphasis on fluency can sometimes lead to the decline of accuracy in speech” (2010: 144). For this reason, imitative speaking tasks can be valuable in promoting accuracy in pronunciation. Their level of focus can range from a word to the whole sentence, depending on which phonological criterion the construct has been defined (see Example 3 below).

Example 3: Word repetition task

Test takers hear: Repeat after me

“sheep” [pause] “ship” [pause]

“sweep” [pause] “swept” [pause]

“The rain in Maine stays mainly in the plains.”

Test takers repeat the stimulus

The stimulus can be delivered through an aural or visual channel, or in other words, it can be spoken or written. In the case of a written stimulus, test takers are instructed to read aloud. In this case, the dependence on memory is minimized, but the task itself is less authentic in terms of both situational and interactional authenticity.

Intensive speaking tasks are also known as Controlled tasks, since they are cued in such way that test takers have a limited number of possibilities when responding to a task prompt. Their purpose is to elicit responses targeting expected language forms, e.g. antonyms or

particular grammatical forms. Intensive speaking tasks may require limited oral production (in Bachman's sense, 1990) or production of stretches of speaking structured by task cues. The following are examples of intensive speaking tasks:

- directed response tasks
- read aloud tasks
- picture-cued tasks
- map-cued tasks (giving directions)
- sentence/dialogue completion tasks
- oral questionnaires
- oral translation of limited stretches of discourse (word, phrase, sentence)

Responsive speaking tasks shift toward more open-ended and less structured response, allowing test takers more freedom of choice regarding both grammar and vocabulary. Responsive tasks require two persons at minimum, each with their own role and communicative goals. The following are examples of responsive task types:

- question and answer
 - questions eliciting structured responses
 - questions eliciting open-ended responses
 - tasks prompting test takers to ask questions (see Example 3.4 below):

Example 3.4: Asking questions

Test takers hear:

- a) *If you could interview your favorite actor/actress, what would you ask them?*
- b) *Ask me about my hobbies or my favorite travel destinations.*

- giving instructions and directions
- paraphrasing (written or spoken input)

Interactive speaking tasks involve relatively longer stretches of spoken output than those discussed above. Another characteristic of interactive test task types is the amount of interaction between interlocutors participating in the execution of tasks. For this reason, interactive tasks can be described as interpersonal (Brown and Abeywickrama, 2010:167), since the interaction in

question involves contributions coming from all parties involved in the task. It is important to note that all task types in this category may be *formal* or *informal*, *summative* or *formative*, depending on the purpose of assessment or the instructor's intentions (for more on assessment techniques, see Brown and Abeywickrama, 2010). The following are examples of interactive speaking tasks:

- interview
- role-play
- discussions and conversations (both as formal and informal assessment techniques)
- games (usually informal and formative)

Extensive speaking task types are tasks that involve long stretches of oral production (extended production in Bachman's sense, 1990). They are somewhat similar to interactive speaking tasks, in that that they can be complex and offer test takers freedom to be creative in using the language. However, the amount of interaction among interlocutors is reduced with these tasks, as they are mainly transactional (*ibid.*). The following are examples of extensive speaking tasks:

- oral presentations
- picture-cued story-telling
- retelling a story/news (from written or spoken input)
- oral translation (of extended prose or technical vocabulary text)

The examples of speaking tasks above are adapted from Brown and Abeywickrama's *Language Assessment*(Chapter 7, 2010); and they are but a selection of possible test task types that test developers may include in speaking assessments. Although they are by no means exhaustive and final, they pinpoint the mainstream tendencies in oral assessments. It should be noted that test developers make the actual selection of test tasks depending on the purpose of assessment and intended use of test results, while at the same time, they bear in mind the qualities of good assessment practice (test usefulness).

4 Authenticity in language assessment

4.1 What is authenticity?

Communicative language ability and theoretical models based on it have been the subjects of research ever since Hymes's *Theory of language use in social life* was published in 1971. According to Hymes, language learning proves its value in the real life, that is, outside classroom, when learners engage in communication with other people (Hymes, 1971). This led to major changes in language classroom activities, which gradually started incorporating new ideas, accompanied by different kinds of more or less authentic learning material and communication exercises (Luoma, 2004). Nowadays, decades since it came into use, this approach to language teaching, and eventually to language testing, still calls for extensive research and clarification. So, what is authenticity?

According to Douglas, authenticity is so important that is identified as a “central concept in specific purpose language testing” (Douglas, 2000: 114). In linguistic practice, however, it cannot be said that there is a universal consensus as to what constitutes authenticity in language teaching and assessment, but what majority of researchers agree on is that authenticity relates to how language is used in non-pedagogic, natural, and non-test communication. When it comes to authentic assessment, Mueller states that it can be regarded as the “measurement of the degree to which students can apply classroom learning to experiences beyond classroom”, because students are asked to perform tasks which they are likely to encounter in the real life, and by doing so, they are expected to demonstrate knowledge, skills and abilities that matter outside classroom as well (Mueller 2005 in Zilvinskis, 2015:7). However, both these are very general claims, which are of little use to test developers and test users who need to operationalize test constructs so as to allow test takers' language ability to be engaged by responding to test tasks.

Authenticity in classroom settings is often analyzed in terms of materials used as stimulus material for eliciting test takers' responses. What test takers normally respond to are spoken and written texts used as input or stimulus material, as well as the test rubric created to set the task context. Referring to *authentic* texts, Morrow refers to the texts containing “a stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort” (1977:13). Losada et al. identify authentic texts as those which

have undergone no teacher's or assessor's intervention, but are presented to learners/test takers in their original form (2017: 92). Using such materials with the goal of enhancing authenticity comes with both support and criticism from experts on language learning and testing. Peacock argues that authentic materials increase students' motivation and "concentration on the task" (1977:152). Harmer (1994) refers to authentic materials as a means of "helping students improve their language production, acquiring the language in an easier manner, and increasing their confidence when using the language in real life situation" (in Losada et al., 2017: 92). On the other hand, there is a lot of criticism related to using the materials which have not been intervened on to better suit the language level of students/test takers. Al Azri and Al-Rashdi claim that weak learners feel "frustrated" when confronted with authentic materials which are above their level, since they do not know adequate vocabulary necessary to process authentic texts (ibid.).

Widdowson tried to explain authenticity in terms of uses that spoken and written texts are put to, rather than the texts themselves. For better understanding of this notion, Widdowson suggests making a difference between two terms – *genuine* and *authentic*. The former refers to actual spoken or written texts produced by language users; the latter, on the other hand, refers to activities and processes related to language use (Widdowson, 1979, 1983). In other words, language testers can decide to use a genuine stimulus material as a prompt for a test task, but if the test task itself does not engage test takers' language ability in an authentic way (the way that language task would in a target language use situation), the interaction between the task characteristics and test takers' language ability would not be authentic. To illustrate this, we can imagine a testing situation when learners are presented with a genuine text in the prompt, such as an actual bus timetable, and the asked to write a paragraph on the frequency of buses on a certain route. Despite the text being genuine, it does not call for an authentic reaction between the input and expected response. People check a timetable and then make an inquiry about the price of the ticket, or the ticket validity in the case of a return ticket, and they can do that in both speech and writing, but they do not normally write a paragraph on the route frequency. However, the same testing situation can yield an authentic response to the stimulus. We can take the same situation, but in the context of vocational training for bus dispatchers and Controllers. Such posts may require submitting reports regarding bus service on certain routes, and in this context, the task of

writing a paragraph regarding the service frequency on a given route will yield an authentic response in addition to a genuine text.

4.2 Situational vs. interactional authenticity

The distinction between genuine and authentic helped the research that emerged in the 1990s, when Bachman further developed the idea of understanding authenticity as the interaction between a language user and a text, or more precisely as a function of an interaction between a language user and a discourse (1991). The most important contribution to understanding authenticity came after Bachman distinguished between two kinds of authenticity (ibid):

- situational, and
- interactional.

Situational authenticity refers to important characteristics of language tasks identified within a particular target language use situation. Once a target language use situation has been identified within the target language use domain (for detailed discussion of domains and TLUs see Chapter 4.1), this situation undergoes a thorough analysis, based on a checklist, or a Framework of task characteristics (or facets, the term used by Bachman and Palmer, 1996). The results of the analysis are then used as a basis for test task characteristics, which, consequently, share the same characteristics as the TLU language tasks. This goes to show the importance of the relationship between the language tasks in the target language use situation and the language test in an assessment (Figure 4.1).

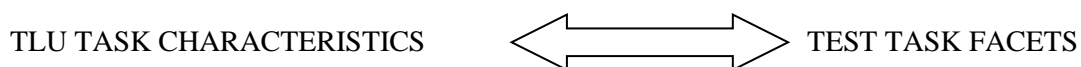


Figure 4.1: The correlation between TLU task characteristics and test task characteristics

It is in specific purpose language tests that this relationship plays an essential role, because the inferences made about test takers' language ability are used for making important decisions about their future. For example, based on test results, test takers may be considered for hiring by an employer, and it is the employer who needs to be persuaded that test takers really possess language skills required by the post. To ensure a high degree of situational authenticity, Douglas proposes that test developers should take two important steps (2000):

- (1) describe a target language use situation, specifying the context and language task features, and
- (2) specify how these features can be implemented through the process of test development, so that test tasks bear resemblance and contain the characteristics of the TLU tasks in order to engage test takers' language ability.

While the first step hinges on the need to emulate characteristics of the context, the second step reflects the need to enable the interaction between the characteristic of the real life tasks, e.g. by providing authentic texts to be processed while responding to test tasks, and the test takers' language ability. This is of utmost importance for the contemporary understanding of authenticity as something that does not exist outside a learner, i.e. in the input or task prompt. In other words, authenticity "resides in language users as they interact with texts and tasks" (Douglas, 2000: 114). The steps described above are built on Bachman and Palmer's recommendations for designing authentic test tasks, where they suggest that the first step should be made by using the framework of test task characteristics so that the critical features that define TLU tasks can be identified. The subsequent step is related to designing test tasks which employ the features identified by the test task characteristics framework (1996). Regarding the authentic material which test takers process while responding to test task, Douglas makes a clear distinction between *input data* and a *prompt*. The former refers to genuine material imported from target language use situation, whereas the latter refers to the intervention on the part of test developers to set up a specific purpose situation in the test itself. In considering the authenticity of input data, test developers need to be aware that the material used should meet the requirements of both situational and interactional authenticity. In other words, simple simulation of real life texts and tasks (situational authenticity) is not enough if the characteristics of such texts and tasks fail to engage test takers' language ability. Authenticity is best achieved when target language use tasks are analyzed well enough so that their properties can be transferred to the test tasks by the means of test rubric, prompts, and input data, provided they are all defined well enough to ensure authentic interaction of the task and the test takers' specific purpose language ability (ibid.).

4.3 Critical elements of authentic assessments

The importance of authenticity and more specifically authentic assessment is probably best reflected in terms of Bachman and Palmer's test usefulness, since authenticity is one of the indicators of test usefulness. As such, it is closely related to the notion of construct validity, proving not only that the test measures what it is intended to measure, but also proving that the inferences based on the test results are valid and can be utilized by test users. Ashford-Rowe et al. (2014) identify eight critical elements of authentic assessment, which need to be considered in the process of test development:

- Challenge
- Outcome: performance or product
- Transfer of knowledge
- Metacognition
- Accuracy
- Environment and tools
- Feedback
- Collaboration

Authentic assessment has to ensure that test tasks are **challenging**, since the real life tasks contain a degree of challenge forcing speakers to produce and construct the meaning and knowledge, rather than to simply reproduce the meaning and knowledge created by others (Newmann, Marks, & Gamoran, 1996). Challenge should not be confused with unreasonably demanding tasks, especially in the light of classroom assessment. Brown and Abeywickrama suggest that tests should be 'biased for best', or in other words they should be designed in such manner that test takers can demonstrate their best performance on them (2010:44). This is an important consideration since authentic assessment strives at linking classroom activities to the real life. Further to this, test takers need to be exposed to a wide assortment of challenging tasks and activities, if inferences about their ability are to be valid.

The **outcome** of an authentic assessment takes the form of either a product or a performance. The reason for this lies in the simple fact that outside classroom, test takers need to demonstrate that they are able to do or make something. In the linguistic sense of the word,

performance and production are inseparable from the language use, especially in a work environment. It is on the part of test developers to determine the knowledge, skills and abilities that are essential for crafting the outcome in the form of a product or a performance, ensuring that this outcome has relevance outside the testing context as well (Brown & Craig, *Assessment of Authentic Learning*, 2004).

Another link to the real life domains can be formulated as the requirement for the authentic assessment to ensure **the transfer of knowledge**. In other words, the skills, knowledge and abilities being assessed should prove to be valuable outside a single content area (Ashford-Rowe & Herrington, 2014). Outside assessment contexts, knowledge is often drawn from a number of different content areas and a range of domains. By supporting the “notion that knowledge and skills learnt in one area can be applied within other, often unrelated areas” authentic assessment proves that it is relevant (Berlak, 1992 in Ashford-Rowe & Herrington, 2014: 208). With regards to its ability to endorse the transfer of knowledge, authentic assessments enable test takers to apply the knowledge and skills across domains, so as to prove the relevance of the tested content to non-pedagogic contexts.

Metacognition in authentic assessment refers to the process of self-evaluation (or self-assessment) and critical reflection of one’s own performance (Ashford-Rowe & Herrington, 2014). This performance is broadly understood as an achievement resulting from a certain action; however, in linguistic sense, it refers to a performance involving the use of a language learnt. Custer notes that “monitoring their own learning through *self-evaluation* can enhance student learning” (Custer, 2000: 29). This means that once students become aware of criteria for correctness and they learn how to apply these criteria in assessing their own performance, they become more independent and take a larger portion of responsibility for learning and progress. Klenowski (1995) defines self-assessment as “the evaluation or judgement of ‘the worth’ of one’s strengths and weaknesses with a view to improving one’s learning outcomes (in Ross, 2006:1). Ross argues in support of using self-assessment in classroom assessment, pointing out the following findings suggesting that self-assessment:

- produces consistent results across items and tasks,
- provides information about student achievement that partially corresponds to the information gathered by teacher assessment,
- contributes to higher student achievement and improved behavior (ibid.)

Self-evaluation, as a form of awareness of one's own performance and learning can complement *peer-evaluation* (or *peer-assessment*), or critical awareness of the performance as demonstrated by peers. Topping (2007) defines peer-assessment as "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" (in Kearney, 2013; 879). As Luoma suggests, this is a promising area for learning as it helps students achieve three important goals: (1) they stay focused on the activity taking place in their classroom; (2) they are aware of their own learning goals; (3) they learn from each other (2004: 189). In all three cases the focus is on learning, but Luoma suggests that peer-evaluation is quite useful in speaking assessment, where students do not have to be responsible for rating linguistic performance of others (although this is not explicitly excluded), but they can be equally qualified judges of task-related performance (for example, they can learn how to assess whether the task has been completed successfully or not). However, for this process to be successful, rating criteria need to be modified so that learners understand what descriptors mean and how they can be used.

In a professional setting, the ability to demonstrate initiativeness and independence in taking actions and making decisions is crucial for keeping the job and being promoted. In educational setting, it helps increasing the overall understanding of the learning process. Given the importance of metacognition, Ashford-Rowe & Herrington believe that it "stimulates deep learning" (2014: 208), and the knowledge stored in long-term memory can be applied over and over in the same or in different domains, at the same time ensuring an effective transfer of knowledge.

The requirement of **accuracy** refers to the value that assessment has to the real life application, especially in the context of work environment. More specifically, students should see the benefits of the assessment, provided it tests the knowledge and abilities that address the needs of the real work environment. To do this, the assessment itself should provide close links between the task and the conditions under which it is carried out and assessed, and in this way simulate and measure the ability in the way it is demonstrated and measured in the real life (Herrington & Herrington, 2006). This and the next requirement link Ashford-Rowe et al.'s approach to that of Bachman and Palmer's, where the latter suggest using Test task characteristics framework to link test tasks to the real life tasks.

The following element of authentic assessment is helping test developers in considering “the fidelity of the **environment** within which the assessment is to occur” (Ashford-Rowe & Herrington, 2014: 209), as well as the **tools** required to complete the task. Given the complexity of the environment in which tasks are carried out, here it would be useful to take into consideration Bachman and Palmer’s Framework which helps recreate the environment in terms of the physical characteristics of the setting, as well as the characteristics of test takers responding to the task. The tools discussed here refer to a wide range of cultural elements, including, language, visuals, and topics linking the TLU task to the test task.

The following element, that of a **feedback**, is considered to be critical at a workplace, where it normally occurs in two ways, as taken and given. The ability to discuss, receive and give feedback is what stimulates improvement and critical understanding of one’s own performance, as well as the performance of others. Due to this link to the real life situations, Ashford-Rowe et al. claim that the opportunity of giving/taking feedback should be built in the test design, in a particular assessment (2014). This idea of classroom activities being connected to the real life, is not entirely new. Namely, when they came up with *Five Standards of Authentic Instruction*, Newman and Wehlage recognize that instruction proves valuable to students if what they are learning is applicable beyond the boundaries of their classroom (1993). In the same vein, authentic assessment should contain elements, one of which is the possibility of taken/given feedback, which will bridge the gap between academic performance and its applicability at a workplace. At the same time, feedback is a means of ensuring that the assessment activity equips students with “interpersonal skills, logic and rhetoric” necessary both in pedagogic and non-pedagogic settings because it can help them determine areas of improvement, and that is “the key to progress” (Ashford-Rowe & Herrington, 2014:210). Brown and Abeywickrama state that grades and scores alone “reduce the linguistic and cognitive performance data available to student to almost nothing” (2000:39). If they are to be meaningful, they should be accompanied by comments and feedback, fostering future learning and revisiting personal as well as course goal and objectives. Luoma states that a useful feedback is “concrete and descriptive, and it relates examinee performances to goals” (2004: 189). If it were to be successful, learners should get a clear picture of what went well and what needs improvement so that they can act upon it accordingly. Luoma suggests that structured feedback mechanisms need to be developed so that feedback can be given or taken meaningfully. She proposes either using

rating checklists for developing structured feedback or “organizing feedback reports in terms of course learning goals” (ibid.).

The value of **collaboration** is recognized as one of the eight elements of authentic assessment, again because collaboration is indispensable quality at a workplace, and as such it is to be stimulated both in formal and informal forms of assessment. Lebow and Wager suggest that the importance of collaboration lies in the opportunities that it gives to educators to engage students in authentic activities which “(a) shift from all students learning the same things to different students learning different things; (b) create group problem-solving situations that give students responsibility for contributing to each other’s learning and (c) help students see the value of what they are learning and choose to share” (in Ashford-Rowe & Herrington, 2014:210). In contemporary teaching methodology, the value of collaboration is often well recognized in a number of activities incorporated in instruction, learning and assessment. For example, the requirement to demonstrate the ability to work in a team is more than obvious at a workplace, however, it is only recently that classroom activities started focusing on shared responsibility for the outcome, in terms of dividing students into groups (teams) and insisting on their joint efforts to complete the activity successfully. Spoken production, in the form of group presentations is easily assessed, by using rating scales which include both individual contribution and the group performance, and in this manner spoken assessment acknowledge the value of collaboration, while, at the same time, they prepare students for modern work environments, where collaboration is no longer a requirement but a must.

5 ESP target language speaking tasks and test tasks

5.1. Target language use domains and target language use situations

This chapter offers an overview of target language use domains, describing their relevance to specific purpose language testing. Additionally, the chapters provides the reader with insight in the hierarchy existing between target language use domains, situations and tasks, explaining the procedures that language testers apply when identifying the target language use tasks. The latter is of utmost important to the research presented in this thesis.

5.1.1 Describing a target language use domain

Communicative language use takes place in various situations and for various purposes, but what they all have in common is the participants' intention to realize their communication goals for which they engage in communication in the first place. Ideally, a test situation should correlate with a language situation outside the testing context itself, keeping in mind that certain goals must be achieved in order to prove that language takers actually possess the skills and abilities they can use in the real life contexts. The Common European Framework of Reference for languages defines a context as a “the constellation of events and situational factors [...], both internal and external to a person, in which acts of communication are embedded” (Council of Europe, 2001:9). Douglas defines a context as a series of external and internal factors determining the direction of communicative acts (Douglas, 2000:43), and to an extent, he uses this term in the same way as a target language use domain, like some other authors do (Bachman, 1990: 342; Bachman and Palmer, 1996: 102).

5.1.2 Target language use domain vs. target language use situations vs. communicative language goals

There is a certain hierarchy when it comes to language use domains and language situations taking place within them. Domains are superior to language use situations, meaning that a number of various language use situations occur within corresponding language use domains. Following the same logic, communicative language goals are inferior to language use situations. In other words, individuals approach language use situations with intentions to achieve their own communicative goals by taking into consideration their language knowledge,

strategic competence enabling them to put the knowledge into use and affective attitudes towards the given situation and their own role in it.

Generally, target language domains can be divided into different categories. The CEFR recognizes four different domains: *personal*, *public*, *occupational*, and *educational*. Within each of these domains, language use is set in the contexts of various target language use situations (COE, 2001:45). Bachman and Palmer, for example, divide target language use domains into two broad categories: educational domain and real life domain (1996: 44). They regard the target language use domain as a “series of specific language tasks which take place outside a language test”, whereas the purpose of the test is to engage test takers’ language competence necessary to solve the corresponding real life language tasks. Having selected the appropriate target language use domain, test takers proceed by identifying language tasks within corresponding target language use situations in order to create language test tasks with the same construct. What follows is a discussion on how to carry out the selection of tasks, but at this point it is worth mentioning that most researchers agree on using a test task characteristic framework as a sort of a checklist to help test developers in representing the construct as fully as possible. One of the most commonly used test task characteristics framework is the one developed by Bachman in 1990(later modified by various researchers, such as Bachman and Palmer, 1996:47; Douglas adapted the framework so as to suite the purposes of testing language for specific purposes, 2001:51, Milanovic further modified the framework within the context of computer-assisted language assessment, 2010:5). Many other researchers in language assessment find this framework to be quite useful when identifying test task characteristics and determining the extent to which they correspond to language tasks in target language use situations of interest (Bachman and Palmer, 1996; Alderson, 2000; Read, 2000; Buck, 2001; Douglas, 2001; Weigle, 2002; Luoma, 2004; Purpura, 2004; Chapelle and Douglas, 2006; Milanovic, 2010).

5.1.2.1 Identifying target language use tasks

Language test developers may be more or less familiar with target language use domain, but it is hard to imagine them being familiar with every single situation taking place within that domain. Given that the success of testing process depends on how well the construct is represented through test tasks, it becomes clear why the familiarity with target language situation, and tasks taking place within it, play a pivotal role in the process of test design. In

literature on language assessment, the following procedure is devised in order to tackle this problem: (1) seek help from field-specific experts who will provide insight into potential language use situations and language tasks that can be performed in it; (2) make a plan for collecting data related to important pieces of information pertaining to the language tasks that can be identified within the target language use situations; (3) devise a data collection plan in collaboration with the given experts; (4) analyze language task characteristics by using a test task characteristic framework; (5) group tasks based on the similar characteristics identified by the framework (Bachman and Palmer, 1996: 102). One of the most enigmatic fields in the real life domain is that related to specific purpose language use. For example, a test user may require that test developers create an assessment tool aimed at investigating language competence of prospective candidates seeking employment at commercial ocean liners. It is hard to assume that any test developers can consider all the possible target language situations and language tasks that they include. Facing a problem such as this one, Douglas suggests that test developers should follow the following procedure: (1) describe and analyze a situation from the perspective of language users in a given situation (what he describes a “grounded ethnography” approach; (2) investigate how native speakers use the language in a given context or target language use situation (“context-based approach”), and (3) seek help from field-specialist informants (“specialist approach”). All these approaches encompass detailed quantitative and qualitative analyses aimed at providing valuable insight in target language use situations and language tasks that can be performed with the purpose of achieving various communicative goals (for more see Douglas, 2000: 93-99).

5.1.3 Construct definition

Once they have identified language tasks of interest, test developers proceed to the following step – construct definition. This step is an essential one since it pinpoints language competences, or their components, which are crucial for successful completion of the task and achievement of communicative goals. Green argues that all tests involve constructs, but it is on test developers to “define, describe and justify the knowledge, skills or abilities they intend to assess” (Green, 2014: 173). Below the author will define a construct as a term, and provide a brief review of the literature regarding the way test developers can define and operationalize

constructs in language assessments. Additionally, this thesis will endeavor to explain what construct definition may include and how it may depend on the type assessment it is used in.

5.1.3.1 *Constructs*

The field of language assessment emerged under the influence of psychometrics following the same basic principle – a competence can be measured, as well as some other psychological characteristics such as intelligence or personality. Considering the problem of validation in psychological measurement, Mesick defined measurement as the process aimed at determining “how much of something there is in an individual”, adding that the starting point would be to determine the nature of “that something” (Mesick, 1975:957). In line with this claim, if we want to use tests as measuring instruments, we need to determine what it is that we want to measure in the first place (Bachman, 1990: 255). In psychology, the “thing” measured is called construct and it refers to latent traits which cannot be measured directly, but rather indirectly based on the manifestation of the behavior that these characteristics cause (Fajgelj, 2009: 315). Brown and Abeywickrama consider constructs to be embedded in every theory, hypothesis or a model that “aims at describing perceptible phenomena” (Brown and Abeywickrama, 2010: 33). Examples of language constructs are *fluency* or *communicative competence*, which *per se* are abstract and cannot be measured or observed directly. Alderson et al. underline that a construct should be understood as a psychological concept given its abstract nature (Alderson et al., 1995: 17); even Alderson himself insists that “a construct is not a real entity” but an abstraction which is defined in accordance with a particular test purpose (Alderson, 2000: 118). When it comes to language assessment, however, it is based on a premise that constructs such as those mentioned above can be measured, because test tasks are instruments helping learners make their latent trait “visible”, and hence measurable. In line with this idea, language tests can be regarded as an operationalization of theoretical construct definitions, the most important purpose of which is to ensure that test tasks engage test takers’ characteristics (their language knowledge, strategic competence, affective characteristics, and background knowledge), and it is through this interaction that an abstract construct definition comes to life, becoming visible and measurable. The most important question to start with is: how to define a construct in a particular assessment?

5.1.3.2 How to define a test construct?

Construct definition relies on the purpose of assessment and intended use of test results, and for this reason there are different possibilities when it comes to formulating test construct definition. Bachman and Palmer differentiate between construct definition based on a course syllabus and the one based on a theoretical model of language competence (1996: 117-118); Douglas, on the other hand, claims that background knowledge (which Bachman and Palmer refer to as “topical knowledge”) can also be a part of the construct measured, so for this reason it must be included in the construct definition (Douglas, 2000); Alderson, however, warns against this practice since background knowledge is often a source of a construct-irrelevant variance in assessing language knowledge, and may cause the so called “method effect” and contaminate test scores (Alderson, 2000:123). If a construct is defined based on a course syllabus (hence the term **syllabus-based construct definition**), its intended purpose is to provide information about students’ strengths and weaknesses, or how well learning objectives have been achieved within a specific educational domain. On the other hand, the purpose of assessment may have nothing to do with the context of education, but a person’s language knowledge need to be assessed, so that test scores and inferences based on the scores can inform decisions related to employment, or immigration. In such cases, test developers cannot write test items following any particular course syllabus, so that they “have to rely on a theory or a model of language ability”, or more specifically on its components describing a particular (language) behavior that is to be assessed (Bachman and Palmer, 1996: 117). In this case we talk about **theory-based construct definition** which is applied in the real life domain. However, there are situations when even within an educational domain, test constructs cannot be defined based on the syllabus, so that a theory-based construct definition should be employed. For example, when students are enrolled in a course, based on a proficiency test results, the test construct is usually defined on the basis of a theoretical model, since there is no particular course or a syllabus to inform test developers’ decisions (Weigle, 2002). Later on, students are usually presented with various forms of formative and summative assessments, the aim of which is to facilitate learning (in the case of the former) or to measure progress or achievement (in the case of the latter). All these

considerations mentioned above suggest that assessment is an iterative process, with numerous decisions and actions calling for revision and improvement.

5.1.3.3 Construct components

Defining test constructs involves specifying the exact components of a language ability that will be observed and measured by an assessment. According to Bachman and Palmer, many language testers are on the wrong track, thinking about language ability from a “unitary, holistic” perspective (1996: 131); instead, they advocate the componential approach, according to which language ability involves language knowledge, strategic competence, and topical knowledge (Bachman, 1990: 84; Bachman and Palmer, 1996: 116-117). In other words, depending on the purpose of assessment, test developers focus on different components of language ability in their attempt to define a test construct and develop test tasks that will operationalize that construct, depending on what inferences need to be made based on test results. For example, if a test is developed for classroom use, and the intended use of test results is to check students’ progress regarding grammar knowledge, test construct will, inevitably, focus on language knowledge rather than other components of language ability – strategic competence and background knowledge. If, on the other hand, the purpose of an assessment is to inform employment decisions, the test construct may include all components of the language ability giving those separate weightages so that test users can select candidates based on the job requirements. When it comes to making employment-related decisions, background knowledge is often given precedence to language knowledge. In some other cases, background knowledge can be taken for granted, so the only decision that matters is the one regarding language knowledge. For example, if selection decisions need to be made regarding hiring mechanical engineers holding a PhD degree to teach engineering at a university where English is the language of instruction, these decisions may be based solely on the language knowledge, provided English is not the candidates’ mother tongue.

The process of test construction involves many steps, which depend on one another, emphasizing the iterative nature of test development process. However, the step of defining the construct may be relatively more important than the others are, since it may affect the validity of test results and inferences based on them. Additionally, the construct definition will influence

test developers' decisions regarding the tasks, since test tasks are the way to operationalize the construct; or, in other words, to make it work. These two steps need to be carefully revised in order to avoid two common traps alluring test developers in their effort to operationalize the construct: *construct-underrepresentation* and *construct-irrelevant variance* (Messick, 1989, in Buck, 2001). The former refers to incompleteness in addressing all parts of the construct by the input materials and corresponding tasks which should engage test takers' language ability and enable the interaction between the ability and test tasks. The latter is the case of putting validity at risk by requiring that test takers demonstrate skills and abilities which have not been defined by the construct. For example, in computer-assisted language test, computer literacy or the lack hereof may interfere with actually responding to test task (in this case with listening to a passage and responding to test questions), which will consequently affect test scores and finally inferences made on the basis of the scores. In other words, the test takers listening skills may fail to be engaged because of their inability to manipulate computer equipment, or because of the equipment's deficiency in certain test administrations (Milanović and Milanović, 2011).

Given the importance of the intended use of test scores, and their correlation with the purpose of assessment and the corresponding construct definition, in the chapter below, the author will analyze possible patterns of correlation among the three.

5.2 ESP speaking tasks

The purpose of an assessment and the intended use of assessment results are key considerations in the process of test task development. In the context of speaking assessment, tasks will involve activities taken by speakers who use language “for the purpose of achieving a particular goal or objective in a particular speaking situation” (Bachman and Palmer 1996 in Luoma, 2004: 31). With these considerations in mind, test developers design language test tasks, which, optimally, address the ability to be assessed, and “cover” the construct as completely as possible in order to provide solid foundations for making inferences on a candidate’s ability to use the language in target language use domains. Target language use domains are of particular importance in communicative language testing, as they cover a multitude of situations where candidates are supposed to demonstrate their proficiency in a foreign language. In learning-based assessments, however, tasks are prevalently based on syllabi allowing test results to show the progress of students. Whatever the context of testing speaking, Luoma argues that test designers need to create instructions both to test takers and to interlocutors/examiners, the tasks themselves, including the materials used throughout the test administration, for example pictures, graphs, role play cards, and other types of stimulus materials (Luoma, 2004: 29). Designing test tasks includes making decisions related to the following key considerations: stand-alone or integrated testing, testing micro and/or macro skills, construct-based or task-based approach, live or tape-based (or recorded) test mode, question format, stimulus material, etc. (for more on considerations in designing test tasks, see Bachman and Palmer, 1996; Luoma 2004, and Brown and Abeywickrama, 2010). Building on the previous research in test task characteristics, more specifically on the work done by Bachman and Palmer, and modifications suggested by Luoma referring to speaking assessment, as well as the guidelines provided by Douglas to address the needs of special purpose language assessment, in this dissertation, modified test task characteristics frameworks will be used to analyze TLU speaking tasks and the corresponding speaking test tasks respectively.

5.2.1 TLU task characteristics framework

The primary purpose of language tests is to make inferences that generalize to those specific domains in which test takers are likely to need to use the target language. In other words, we want to make inferences about test takers' ability to use language in a target language use (TLU) domain. For this reason, care should be taken for test tasks to resemble the target language use tasks, and this can be achieved through simulating language task characteristics, thus ensuring *situational authenticity*. Moreover, test tasks should be designed in such a manner that the interaction between the test taker and the task is similar to the interaction between the language user and the target language use situation, or in other words, a care should be taken that the task allows for *interactional authenticity* (Buck, 2001: 108). The way we develop test tasks, i.e. their format, contents, nature, the media in which we present them, the equipment and facilities we use to administer the test, may all influence the test takers' performance on the test, and consequently the inferences made on the basis of that performance. This influence is also known as "test method effect" and is documented in the research on second language testing because it can affect the validity of test scores and, consequently, the validity of inferences based on the scores (see Bachman, 1990). For this reason, it is necessary to take precautionary measures when developing test tasks and carefully analyze their characteristics.

TLU task characteristics are worth considering for several reasons. First, they provide us with an insight of what constitutes language tasks and how they can be linked to test tasks, what links there may exist between these two groups of tasks, enabling us (as test developers) to develop test tasks which correspond to (target) language tasks. Second, test task characteristics will help determine the extent to which a test taker's language ability is engaged. Third, the degree to which test task characteristics correspond to particular target language use task will determine the authenticity of test task as well as "the validity of inferences made on the basis of test performance" (Bachman and Palmer, 1996:43). The framework of TLU task characteristics that will be used in its modified form (described below) builds on those originally proposed by Bachman (1990), Bachman and Palmer (1996), with modifications suggested by Douglas (2000), Chapelle and Douglas (2006), and Luoma (2004).

Bachman and Palmer developed a framework of language task characteristics, stating that the purpose of their framework was to provide a basis for test development and use. They use the

term ‘task’ to refer to both TLU tasks and test tasks, because they find that the characteristics described in their framework apply to both TLU tasks and language test tasks. There are five aspects of tasks that they set out to describe using the framework: setting, test rubric, input, expected response, and relationship between input and response (1996). The characteristics of test tasks in the framework proposed by Douglas include the following: rubric, input, the interaction between input and response, and assessment criteria. His main intention was to outline a framework of task characteristics in language use situations and language for specific purposes tests that will allow test developers to analyze TLU situation and to develop test tasks which will reflect the characteristics of the target situation. The essential advantage of such framework is that test takers’ performance on the test tasks can now be interpreted as evidence of their ability to perform tasks outside the testing environment (2000). Additionally, Bachman claims that by establishing “a close correspondence between the target language use tasks and test tasks” the test tasks’ authenticity will be increased” (1990: 112).

The framework of TLU task characteristics which will be used in this dissertation is mainly based on the considerations recommended by Douglas and Chapelle and Douglas, whose frameworks include many elements of Bachman and Palmer’s test task characteristics framework. Test task characteristics framework, first developed by Bachman in 1990, was revised by Bachman and Palmer in 1996, but the overall outline remained the same, including the characteristics of: (1) the testing environment (or setting), (2) the rubric, (3) the input, (4) the expected response, and (5) the interaction between input and response (Bachman, 1990; Bachman and Palmer, 1996). Comparing Douglas’s framework to that of Bachman and Palmer’s, one can notice that he made some changes in his framework, so that his includes the characteristics of: (1) the rubric, (2) the input – including the characteristics of the setting in Bachman’s sense, (3) the expected response, (4) the interaction between the input and response, and (5) assessment. Since this thesis deals with assessing the English language for specific purposes, Douglas’s work will be discussed into more detail below.

Building on the work of Bachman and Palmer, Douglas modified their test task characteristics framework, so it can better match the specific purpose language testing. To do so, he advocates deriving test task characteristics from target specific purpose language use situations. As previously mentioned, Douglas differentiates between target language use domains

and target language use situations, saving the term *domain* for discourse domains only. The essential difference between the target specific language use situations and test situations lies in the explicitness of their characteristics. Whereas in the former, the characteristics are often implicit, and embedded in the background knowledge of the participants in a communicative act, in testing situations they have to be made explicit. However, mere simulation of target specific language use situation characteristics does not necessarily guarantee the success of the overall testing process, because some of the characteristics found outside the testing environment are hard to emulate due to various constraints inherent to a testing situation (time, space available, school policy, etc.). In the same vein, the extent to which test task characteristics will correspond to the TLU tasks depends on various factors: the purpose of assessment, the characteristics of test takers, and the resources available for developing and administering the test (p. 49). The reason why Douglas believes that test task characteristics framework is to the benefit of the testing process is the same as the reason why Bachman and Palmer developed their framework in the first place – it allows test developers to analyze a TLU situation and to develop test tasks which reflect the characteristics of the tasks in the target situation. Furthermore, Douglas emphasizes the frameworks applicability in analyzing both TLU tasks and test tasks by using the same set of characteristics (Table 5.1).

Table 5.1: Overview of target language use and test task characteristics, Douglas, 2000: 51-52)

Task Characteristics Framework
Characteristics of the rubric
Specification of the objective
Procedures for responding
Structure of the communicative event
Number of tasks
Sequence of tasks
Time allotment
Evaluation
Criteria for correctness
Rating procedures

Characteristics of the input

Prompt

Features of the LSP context

Setting

Participants

Purpose

Form and content

Tone

Language

Norms of interaction

Genre

Problem to be addressed

Input data

Format

Visual

Audio

Vehicle of delivery

Length

Level of authenticity

Situational

Interactional

Characteristics of the expected response

Format

Written

Oral

Physical

Type of response

Selected

Limited production

Extended production

Response content

Nature of language

Background knowledge

Level of authenticity

Situational

Interactional

<p>Characteristics of the interaction between the input and response</p> <p>Reactivity reciprocal ↔ non-reciprocal</p> <p>Scope broad ↔ narrow</p> <p>Directness Dependent upon input ↔ dependent upon background knowledge</p>
<p>Characteristics of the assessment</p> <p>Construct definition</p> <p>Criteria for correctness</p> <p>Rating procedures</p>

5.2.1.1 Characteristics of the rubric

Both Bachman and Palmer’s and Douglas’s frameworks feature the characteristics of test **rubric**. This term requires certain clarification in the light of language assessment, because it refers to “the characteristics that specify how test takers are expected to proceed in taking the test” (Bachman, 1990: 115). In some other testing contexts, the term is usually associated with rating test takers’ performance, and for the purpose of making this distinction clear, in this thesis we will use the term “rating scales” rather than “rating rubrics” when we talk about criteria for rating performance. Other authors, like Luoma and Buck, use the term *instructions*, emphasizing the importance of their being clear to test takers (see Buck, 2001; and Luoma, 2004). Additionally, Buck warns that attention must be paid that the language of instructions is easier than the level of the language in stimulus material, because any misunderstanding related to instructions may lead to construct-irrelevant variance (2001:119). Douglas uses the term rubric to describe the characteristics of the communicative event including the following: objective, procedures for responding, structure, time (available to complete the task), and evaluation.

The **objective** has more or less the same meaning as the purpose of performing the task, because it describes what it is that a language user is trying to achieve by engaging in the task. For example, in a TLU situation, the objective of a speaking task can be to inform customers about the latest promotion, whereas in the LSP (language for specific purposes) testing context,

the objective may be assessing the range of specific purpose vocabulary and politeness norms in addressing customers. **Procedures for responding** are often implicit in a non-test situation, so the language users know how to proceed with carrying out the task. In a testing situation, test takers must be told explicitly how they are expected to respond to the task. For example, in a multiple choice test rubric, the requirement could be to circle 2 or more answers, and not only one as is most often the case, and for this reason, the requirement should be stated explicitly and not left to a chance. **Structure** of tasks in a non-test situation is often obvious and explicit, however, in a test situation, test takers need to know the number of tasks, their relative importance. For example, in a non-test situation, at a workplace, a secretary may be told that she is expected to book plane tickets before she books the hotel accommodation for a business trip, or she will already know that this is the correct order of tasks. In a speaking task, in a test situation, test takers need to be told explicitly that they will role-play making travel arrangements based on the travel times, etc. **Time allotment** characteristics refer to the amount of time test takers have in order to complete the task. Again, in a non-test situation, this will often be implicit, but in a testing situation it should be made explicit and obvious to test takers. It is interesting to notice that Douglas distinguishes between the characteristics of **evaluation**, which are included in the rubric, and the characteristics of **assessment**, which belong to another set of characteristics. The characteristics of evaluation find their place in the rubric because they refer to the information regarding the assessment of the task that is given to test takers. In other words, evaluation characteristics help test takers with understanding how their response will be evaluated by test raters. Later, when we discuss the assessment characteristics, it will become more obvious that evaluation characteristics are nothing but a simplified version of the assessment characteristics, adapted to suit the needs of test takers. To this end, criteria for correctness and the procedures for rating are spelled out so that they become obvious to test takers, providing them with the clear picture of what is considered to be correct, or sufficient, and how it will be rated. Assigning scores to test takers' responses is based on the assumption that certain responses are correct, while others are incorrect, and that they can be scored as such. Making these explicit in a test rubric is important as it helps test takers in allocating time and applying different test-taking strategies. In responding to multiple-choice questions, for example, test takers are usually told to select "the correct answer". Consequently, this implies that there is only one correct answer, whereas all the other answers provided are incorrect. In some other

tasks, test takers may be told to sequence sentences in a summary, implying that there will be only one correct sequence, etc. With respect to the procedures for rating, test takers may be instructed to use the information provided in the reading/listening passage before they proceed to their speaking task, where their background knowledge is of no relevance to providing the correct response. The extent to which the rating procedures are made explicit to test takers is vital to test takers' awareness of what constitutes a sufficient response (Buck, 2001:122). When a prompt elicits an open-ended response, test takers should know how much as well as what is considered adequate.

From what is stated above, it can be concluded that in non-test situations rubric will often be implicit, embedded in a person's background knowledge and familiarity with the communicative situation they are supposed to engage in. On the other hand, even in non-test situations, the characteristics of rubrics may appear in the form of instructions coming from people in charge, telling language users what kind of performance is expected of them. In a test situation, it is highly unlikely that test developers will leave it to test takers to rely on their familiarity with test taking process in order to proceed to the tasks. On the contrary, they will try and provide as much information as possible to ensure that test takers are able to demonstrate their language ability on the task. Otherwise, there is a risk that test takers do not perform well, not because their language proficiency is not at a sufficient level, but because they were affected by the "test method" (for more details on "test method effect" see Bachman, 1990).

5.2.1.2 Characteristics of the input

The following set of characteristics is that of the **input**. Douglas makes it very clear that in specific purpose testing, it is the input that sets the characteristics of a target specific purpose situation within the testing context, allowing test takers' specific purpose language ability to be engaged in solving test tasks. In other words, the input serves as the means by which contextual features are established and controlled, offering test takers a sufficient amount of cues to engage an appropriate discourse domain and respond to tasks as originally envisaged by test developers. It is important to note here that this author makes a clear distinction between the rubric and the input - the former being a set of specific procedures, whereas the latter refers to the material given to test takers to process and respond to. Further to this, there is another important distinction to consider— that between the **prompt** and **input data**. Although it may not always be

very obvious, the prompt refers to contextual information provided to the language user/ test taker helping them engage the appropriate discourse domain while solving the task. More specifically, the prompt covers the following features of the LSP context: *setting, participants, purpose, form and content, tone, language, norms of interaction, genre, and the problem to be addressed*. It should be noted that not always all these characteristics are present in a prompt, because the prompt is provided by test developers for every item/task/test respectively, and it is often left to a developer's judgment to decide how much contextual information will be sufficient for test takers to understand the task and proceed to solving it. The input data, on the other hand, is characterized by its format referring to the authentic aural or visual material coming from the TLU itself. Regarding the *format* of the input data, it can be noted that input data may also refer to an object, including the equipment that a test taker needs to describe or manipulate in order to complete the task. The characteristics of the input data format include the *vehicle of delivery*, referring to the material being delivered as *live* or *reproduced* (recorded). Finally, the *length* of the input data is either constrained in terms of time or the number of words in a spoken/written text. Whether the input material is also authentic is to be determined by analyzing its authenticity characteristics, and by this, I refer both to situational and interactional authenticity covered by the set of characteristics called **the level of authenticity**. It almost goes without saying that in a TLU situation, input data and responses to it are authentic, because they occur in their "natural" setting. However, in a test situation, they are separated from their situational and interactional context, which may result in them potentially losing their authenticity (Widdowson, 1983, in Douglas, 2000). The set of characteristics related to authenticity is aimed at ensuring that the problem of losing authenticity is minimized. "By taking stock of the situational and interactional features that the input data and response in the test share with the target situation" is the way to preserve authenticity in a test (Douglas, 2000:57). However, test developers are often misled by the source of input material, believing that the amount of technical vocabulary is enough to secure authenticity. The text itself may be genuine enough, in Widdowson's sense; however it may prove to be above the test takers' proficiency levels, and as such it will fail to engage their language ability and the appropriate discourse domains. Consequently, test developers need to consider both situational and interactional authenticity, if they are to claim that test tasks are authentic. To this end, the prompt is provided to establish a specific purpose context in case the data alone do not provide enough contextual

cues. Additionally, if authenticity is to be secured, it is often advisable to ask opinion from a subject specialist, especially in terms of determining the degree of specificity of input data, as test developers may not always be aware of this quality.

Finally, when it comes to the sets of characteristics related to the rubric and input, it must be taken into consideration that they sometimes overlap in reality. It is the purpose of the assessment and the specificity of a task that make test developers decide how many contextual cues to provide and in what form exactly.

5.2.1.3 Characteristics of the expected response

The **expected response** is another set of characteristics which can be found both in TLU situations and in testing contexts. The reason for this is simple – the input data’s role is to provide the basis for analysis and processing based on which a response will be provided. In a TLU, participants in a communicative act have certain expectations when it comes to “the characteristics of their respective responses as the discourse evolves” (Bachman and Palmer, 1996: 53). In a testing situation, and as the term suggests, this set of characteristics is related to what test developers expect that test takers should do – use the language being tested, react physically, or both - in response to the input and the prompt they receive in a task. It should be noted, though, that the test taker’s response could be different from the one expected, and this usually happens for two reasons – (1) the lack of language ability, implying the test taker does not possess sufficient knowledge to solve the task, and (2) problems inside the task itself due to a number of reasons – unclear instructions, insufficient number of cues failing to engage the appropriate discourse domain, task difficulty, etc. With regards to this issue, Bachman and Palmer distinguish between “the expected response, which is part of test design, and the actual response, which may or may not be what was intended or expected” (ibid.). Regardless of the possible variations, test developers still expect a certain kind of response which can be described by the following set of characteristics: *format, type, content, and the level of authenticity*.

When it comes to the **format** of the response, it refers to the manner in which the response is produced. It can also be noted that the way the response is produced depends, to a large extent, on the rubric and the prompt. Following the instructions, test takers may respond by providing a *written, oral, or a physical* response. In other words, test takers can respond by

speaking, writing, reacting, demonstrating a procedure, manipulating a tool or equipment, typing their answers on a computer or tablet, writing on the board or on a test paper, or any combination of these. In discussing the **type** of the expected response, Bachman and Palmer reflect on the traditional assessment practices that distinguished between two types of responses: a *selected*, and a *constructed*. In line with such practices, it was usual to develop a multiple-choice item, where test takers have to choose, or select one response from the several provided alternatives. In more recent testing practice, this kind of task can be made additionally challenging so that the test takers have to choose a limited number of alternatives among many more offered (for example, two correct out of five alternatives). Unlike the widespread belief that this kind of format is lacking in situational and interactional authenticity, there are TLU tasks requiring selection. For example, in a highly specific educational setting, pilot trainees are given a multiple choice for manipulating the equipment for flying a plane. Depending on the runway configuration, the weight of the aircraft, the wind, and other variables, they are given a multiple choice in which they have to select the appropriate option for safe landing. On the other hand, the response may be constructed, meaning that the test taker must construct or produce their response to the task. Bachman and Palmer made additional distinction with regards to the length of the produced response, taking a sentence, or an utterance as a unit based on which responses can result in producing a short answer – a word, phrase, or a sentence – and this kind of response is known as a *limited production* response. Additionally to this, the task may require that test takers proceed by producing an *extended response*, in which case their answer takes the form of a longer utterance/sentence, a paragraph or a longer written or spoken text (Bachman and Palmer, 1996). In addition to the selected and constructed response types, Brown and Hudson suggest considering a *personal* response, particularly in alternative assessments, such as portfolios and projects. In case of a personal response type, it is not the test taker's language ability that is in the focus of attention, but their personal relations toward the stimulus or the task. For example, test takers may be asked to reflect on a particular task, their own performance or that of their peers, which is quite useful in classroom settings, when students are asked to evaluate their own performance, or to provide peer-ratings based on the set of criteria, or when they need to provide commentaries on a project or portfolio (Brown and Hudson, 1998).

Another aspect, highly relevant in specific purpose assessment situations, is that of the *content* of the expected response. This set of characteristics includes the *nature of language* and

specific purpose *background knowledge*, both reflecting the construct to be measured. Douglas finds this set of characteristics to be a key aspect of the expected response because it helps ensure that “the response elicits the necessary aspects of specific purpose language ability [...], so that the intended construct may be adequately measured” (Douglas, 2000: 63). The final set of characteristics related to the expected response, in Douglas’s framework, is that of the **level of authenticity**, both *situational* and *interactional*. In order to make inferences about the test takers’ ability in the specific purpose context, it is necessary that both the input and the expected response demonstrate that they are relevant to the target language use situation. Furthermore, they both need to be plausible, not only in terms of situational resemblance, but also in terms of the interaction between the task and test taker’s language ability.

5.2.1.4 Characteristics of the interaction between input and response

The relationship between the input and response can be described through a set of characteristics termed as the **interaction between input and response**, including the following: reactivity, scope, and directness (Bachman and Palmer, 1996; Douglas, 2000). *Reactivity* is a characteristic of showing the extent to which the input can be changed depending on the responses of the language user in a TLU situation, or the test taker in a testing situation. This usually requires more than one participant in spoken communication, because only when there are two interlocutors present can we talk about the communicative exchange. Douglas states that this quality ranges on a continuum between *reciprocal* and *non-reciprocal* (Douglas, 2000: 63), and since it is the continuum that we are talking about, the interaction can be anywhere from highly reciprocal to non-reciprocal. There are two distinguishing features, according to Bachman and Palmer, to identify reciprocal language use and tasks: (1) the presence of feedback, and (2) interaction between two (or more) interlocutors. It seems natural that in a TLU situation, when there is a communicative exchange going on, interlocutors exchange not only utterances, but gestures and facial expressions as well. Consequently, the feedback they are receiving can affect the subsequent reaction, be it verbal or non-verbal. In testing situation, this may not always be the case, so for the sake of clarity, test developers need to consider the issue of feedback and its role in affecting the response. For the sake of providing a better understanding of the relevance of this set of characteristics to the corresponding TLU situations, we may identify three situations demonstrating the reactivity of the expected response. For instance, two interlocutors

may engage in a *highly reciprocal interaction*, because they can instantly provide a full feedback to each other, in terms of nodding, facial expressions and the possibility to ask for additional information or clarification. In a workplace setting, for example, an executive manager can address a large audience of line managers and employees discussing the future strategy of development. In such circumstances, it is impossible to talk to each of the attendees in person, seeing their reactions and asking their opinion. What is possible, however, is to get a limited feedback in terms of sounds expressing approval or disapproval, or the speaker can rely on the facial expressions of some of the attendees in order to get information if his speech requires more clarification or further details. In this example, where limited feedback is provided, we can define the relationship between the input and the expected response as *somewhat reciprocal*. Finally, if an utterance is a recorded message sent via the phone, or instant messaging application, there will be no instant feedback, and consequently, the interaction between the input and expected response will be *non-reciprocal*. Additionally, Bachman and Palmer recognize *adaptive* relationship in computer-adaptive tests, where the subsequent task's difficulty depends on the test taker's response to the previous task. If they answer correctly, they will be presented with a slightly more difficult task, if their answer was incorrect, the subsequent task will be easier. It can be concluded that such tasks do not involve any feedback provided to the test taker, but as Bachman and Palmer notice "they do involve an aspect of interaction, in the sense that their responses affect subsequent input" (1996: 55). However useful adaptive tasks may be for determining the test taker's level of language proficiency, they are of little use to live oral assessments, so this issue will not be pursued in the analysis of speaking tasks in the rest of this discussion.

The **scope** of the relationship between the input and the expected response pertains to the amount or range of the input - including its variety- to be presented to the language user or test taker so that they can process and respond to it. In LSP testing, there is a trend to provide test takers with varied and a relatively long input, although its length and variety will depend on the purpose of the assessment and the construct being measured. Tasks that require that test takers or language users should process a richer input are characterized as *broad scope*. On the other hand, tasks in which test takers and language users have to process a limited input before they respond can be characterized as *narrow scope* tasks. It should be noted that broad scope tasks might not yield an extended production response. In other words, there is no direct relationship between the

two. As mentioned earlier, the purpose of the assessment will determine how much input needs to be processed and the length of the response that will be based on the input. For example, it is not uncommon in the business world that analysts process a large quantity of information coming from different sources when they perform a market analysis. The results based on such analyses can be expressed in a few words expressing the decision in favor of the market in question or the opposite. On the other hand, the information coming from a few graphs could result in a quite comprehensive report, depending on the situation, purpose and the audience for which the report is intended.

The aspect of **directness of the interaction** between the input and the expected response is the one pertaining to the degree to which the expected response depends on the information in the input. With regards to this set of characteristics, it can be said that directness is placed somewhere on the continuum ranging from fairly direct to highly indirect tasks, with many possible options along the continuum. The decision as to the degree of directness can be arbitrary, but it seems relatively easy to identify the extremes in the continuum, whereas other values can be identified as somewhat direct or somewhat indirect. However, the aspect of directness is important in LSP testing, because many tasks will tend to be indirect, requiring that test takers possess certain background, topical knowledge in order to proceed to solving what should essentially be a language task. If we take, for example, a reading task in which test takers have to read about the causes of inflation in order to solve a multiple-choice reading task, it can be concluded that such task involves a fairly direct relationship between the input and the expected response, because no special background knowledge about the causes of inflation is required. Test takers have all the answers in the text itself. However, if test takers are asked to prepare a five minute oral presentation on the causes of inflation, this requires certain background knowledge to enable test takers to speak about this issue with confidence and demonstration of not only the knowledge of the language in which the presentation is to be delivered, but the topical knowledge of the subject matter – in this case inflation. It is evident that for successful performance on this task, test takers need to possess the topical knowledge in order to plausibly attend to the task, so this task would be highly indirect. Another task can be developed, and placed somewhere towards the middle of continuum, if test takers are provided with the input data based on which they can formulate their answers and prepare the presentation. As Douglas observes, the point at issue in LSP testing is to provide test takers with

sufficient contextual cues in order to engage their specific purpose discourse domains so that they can proceed to responding to the tasks in very much the same way the language users in TLU situations would. In language for specific purpose assessment this also involves the engagement of specific purpose background knowledge, on condition enough contextualization is provided in the form of “specific purpose test task characteristics” so that the appropriate discourse domain can be engaged (2000: 67). If these requirements are met, the inferences based on test takers’ performance on the test task can be interpreted as evidence of their specific purpose language ability outside the testing context.

5.2.1.5 Characteristics of the assessment

The set of characteristics of **assessment** is derived following the approach suggested by Douglas, although there are other approaches as well (see Bachman and Palmer, 1996; Alderson, 2000; Buck, 2001). As mentioned above, one of the most significant changes that Douglas made on Bachman and Palmer’s framework was to distinguish evaluation criteria from assessment criteria. This distinction refers to the former being related to the extent to which test takers are informed about the nature of the criteria used to score their responses, while the assessment criteria refer to the same set of criteria and procedures described in more technical terms. These two sets of characteristics target different audiences - the evaluation criteria are aimed at familiarizing test takers with what will constitute an acceptable response, whereas the assessment criteria are the tool used by test developers and test raters. Douglas suggests that this set of characteristics should include the *construct definition*, *criteria for correctness*, and *rating procedures*. All these characteristics are derived by analyzing the specific purpose target language use situation in order to create a set of characteristics which will bring testing situation closer to the target language use situation. In line with this, Douglas makes a distinction between the construct of language ability in TLU and the construct which will be measured in a language test because the real life performance on a given task is so complex that it makes it almost impossible to emulate all the characteristics in testing context. In addition to this, there are many practicality-related constraints which make the whole process more difficult than it seems, such as finances, time, staff, and educational policies. Whatever the limitations, the assessment procedures are still feasible, so Douglas suggests analyzing the characteristics of language in TLU situation in order to define the construct to be measured in a test. Additionally, the criteria

for correctness can also be derived by analyzing the TLU situation, as well as the procedures for implementing the assessment. Both in TLU tasks and in test tasks, there are certain expectations with regards to how the communicative goals are achieved and how this achievement is evaluated. For this reason, both characteristics – assessment criteria and rating procedures – are inherent to the tasks in a TLU and the corresponding tasks in a test. When it comes to identifying assessment criteria, Jacoby, for example, suggests using a framework based on the concept of *indigenous assessment* (1998, in Douglas, 2000). This means, in practice, that test developers should observe the assessment that takes place in the TLU situation, analyze its components as it is being performed by subject specialists in vocational settings and then apply its characteristics in developing a set of assessment criteria for testing purposes. In this way, it is assumed, the assessment criteria applied in the target language use situation and the assessment criteria applied in the corresponding testing situation share the same set of characteristics. If we consider the practicality of such endeavor, it should be noted that this is a time-consuming effort which does not yield universally applicable set of criteria. For example, test developers and assessment specialists can decide to observe oral presentations in an academic conference, trying to come up with the set of criteria they intend to use in assessing extended spoken production. This involves, but is not limited to, a careful study of the interaction that takes place between the speakers, audience, and subject specialists evaluating the presentations and providing feedback to the presenters. However extensive the set of assessment criteria may be developed to suit this purpose, it is fairly hard to claim that it is universally applicable for spoken production in general. Rather, such set of criteria can be of use for oral presentations in academic settings, more specifically, for extended oral production in similar TLU situations – seminars and conferences. As Jacoby and McNamara warn, target language use situations comprise specific and dynamic characteristics that are difficult to repeat in a testing context no matter how situationally authentic it is (1999). Instead, the assessment criteria derived from this TLU situation would have to be adapted and modified for any other testing purpose related to assessing extended oral production in academic settings. To resolve this issue, Douglas insists on analyzing the construct definition as it is the key to understanding the assessment criteria both in TLU and testing situations. The difference lies in the fact that in the TLU, the construct is implicit, as it is often a “part of the professional or vocational culture”, whereas in testing context it must be specified and stated explicitly in order to ensure that it not only reflects the language

use in the TLU, but test developers' understanding of what specific purpose language knowledge entails (2000: 69). Furthermore, if the construct definition reflects the aspects of language use in the target situation, criteria for correctness and rating procedures must be closely related to the construct, making sure that they cover what the construct states the test assesses. In other words, if we assume that the construct definition represents a theoretical statement of what constitutes a communicative ability necessary to carry out a task in TLU, criteria for correctness can be regarded as an operational construct definition. They must represent the construct, covering it fully, so that assessment results really show how much of the ability test taker has. In the same way, rating procedures are equally important to bring the whole process to an end and quantify the criteria in order to provide test results as a meaningful basis for inferences about the test taker's communicative ability.

5.3 Operationalization: developing test specifications and test task specifications

Operationalization is a stage in an assessment process aimed at developing test specifications (often referred to as *blueprint*), test task specifications and actual tests. According to Bachman and Palmer, "these will have been considered in the selecting and describing TLU task types for possible development as test tasks" (1996: 171). In the coming chapters, the author will provide an overview of the existing test specifications models, propose the actual test specifications model that will be used for the purpose of the research outlined in the introduction of the thesis, and discuss rating scales which are of pivotal importance in fighting subjectivity in rating.

5.3.1 Test specifications

Once the context has been analyzed and described in terms of task characteristics resulting from the analysis of the TLU, test developers can proceed to the crucial step in the test development process: developing test specifications. Following the analogy with architectural design, "specifications are the design documents which show us how to construct a building, a machine or a test" (Fulcher, 2010: 127). Bachman and Palmer, for example, use the term *blueprint*, referring to the plan based on which the entire test is constructed. Their understanding of test tasks differs from the "traditional" one, treating test tasks holistically, so that according to

this old approach, the blueprint usually refers to “the table specifying the number and types of items that are to be included in a test” (Bachman and Palmer, 1996: 180). However, test specifications, in the modern sense of the word, are much more than that. In educational settings, test specifications serve as a document or a set of documents with multiple purposes; they can be used as a kind of a blueprint for test developers and item writers; they are often a reference point for validation researchers; and, sometimes specifications are convenient source of information for score users (Douglas, 2000: 109).

The more complex the test, the more developed test specifications are. For some testing purposes, it is often enough to identify the constructs and provide sample items to help item writers in covering the intended constructs. For other purposes, especially if we talk about high-stakes standardized tests, a whole set of different documents is required in order to ensure reliability and validity of results, as well as to create conditions for the standardization of the test administration. Bachman and Palmer’s test blueprint, consists of two parts: (1) the task specifications, and (2) the test structure elements, including the number of parts/tasks, the salience of parts/tasks, the sequence of parts/ tasks, the relative importance of parts/tasks, and the number of tasks/parts (1996: 176). Mislevy et al. developed a plan which can be applied in majority of testing purposes, listing the total of 5 different documents that can be provided: item/task specifications, evidence specifications, test assembly specifications, presentation specifications, and delivery specifications (2003).

Test specifications document can vary in size and complexity depending on the test purpose and the requirements of a testing context. If a test is complex, the specifications document is likely to be complex as well, detailing various aspects of the test itself and the testing situation. According to Green, specifications usually include three elements: a *design statement*, an assessment *blueprint*, and *task and item specifications* (2014: 29). The design statement normally covers the purpose of the assessment, the identification of the people who will be developing and administering the assessment, the identification of test takers and score users, the skill to be measured, etc. The assessment blueprint follows the analogy of a building architecture, covering the content which will be assessed, the methods used, the number and identification of test tasks and sections, the length, and so on. Following the work of Bachman

and Palmer (2010) and Fulcher (2010), Green summarizes the questions that an effective blueprint should provide answers for (Table 5.2).

Table 5.2: Elements of an effective blueprints document (Based on Bachman and Palmer, 2010 and Fulcher 2010, in Green, 2014: 31)

<p>Assessment content</p> <ul style="list-style-type: none"> • How many components (sections of a test, class quizzes, homework assignments, etc.) are included? • What is the function of each component? What knowledge, skills or abilities is it designed to assess? • How are the components and tasks ordered? • How many tasks are included in each component? • How much time is allowed for the assessment? How much for each component? • The assessment will be administered: <ul style="list-style-type: none"> - How (on paper, via computer, by the teacher)? - When (point in the school calendar, day of the week, time of day)? - Where (in class, in a computer lab, in an examination hall)? 	<p>Reporting results</p> <ul style="list-style-type: none"> • How are results reported? • If scores are awarded: <ul style="list-style-type: none"> - How many points are awarded for each component? - How are overall scores calculated? • If scores are not awarded, in what form will results and feedback be provided? • How are any pass/fail grading decisions made?
--	--

Finally, the task and item specifications document lists the tasks and provides the samples of tasks and items, making it explicit what task an items format are considered to be suitable for measuring the intended construct. Depending on the purpose of an assessment, Green points out that there are two ways to go regarding the task and item specifications: (1) creating task specifications as learning objectives, and (2) creating task specifications as a means of “capturing important features of real life language use” (2014: 36).

A language test, as well as any other test, is seen as a measuring device. This device, however, is useful only if it produces valid and consistent measurements, so that test specifications come into play as a means of ensuring that the test measures what it is intended to measure. In other words test specifications can be considered as a part of the “technology required to craft precision instruments that give the same measurement results” (Fulcher, 2010: 129). According to Fulcher, a test is nothing else but a realization of test specifications. To understand how valuable test specifications are to the inferences based on test results, one has to distinguish between a *test form* and *test version*, because these two are often wrongly thought to mean the same. A test form is generated from test specifications, ensuring that the same constructs are measured in parallel test forms, in various test administrations. This critical feature of the test form, that it is parallel to any other test forms based on the same specifications, practically ensures that each form contains roughly the same kind and number of tasks/items measuring the same construct. If different forms of a test were developed from the same specifications, it is reasonable to consider them comparable, and “without comparability of constructs and task characteristics, any demonstration of statistical equivalence will be meaningless” (Bachman and Palmer, 1996: 178). Over time, test specifications evolve into new versions of the test, due to the changes that test designers decide to make on the test itself. The changes are made because test designers learn new things about their test, realizing that some tasks and items do not produce the intended measurement results, so they have to be replaced by new ones. According to Fulcher:

... sometimes we find that features of the instrument produce variability that we did not expect. The sources of variability are researched. If these prove to be part of what we wish to measure – the construct – the test specifications are changed to allow their continued presence in future versions. If they prove to be construct irrelevant they are a source of ‘error’, and the instrument needs to be redesigned to eliminate it. (2010: 134)

Consequently, a new version of the test is made, making it a requirement that all previous versions be discontinued so that new forms of this latest version can be administered. All the new forms based on the latest version of the test will be parallel, measuring the same construct, as intended by the test designers and test purpose (Figure 5.1 below). However, it is only test versions that are changed, while the test remains the same, measuring the same skill or ability (ibid.).

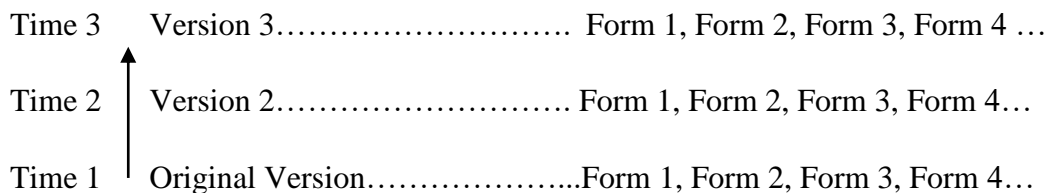


Fig. 5.1: Adapted from Fulcher, Figure 5: Forms and versions (2010: 130)

Together, the specifications discussed above allow test developers to build and deliver a test form for a particular administration, and further to this, test specifications (or the blueprint) play an important role in determining the authenticity of the test. If we think of authenticity as of the correspondence of the characteristics of the test tasks to those of the real life tasks in the target language use situations (in Bachman and Palmer’s sense), the specifications provide a detailed description of the test and the tasks, facilitating the evaluation of the aforementioned correspondence.

5.3.1.1 Test specifications models

Test specifications are often considered to be essential to the process of test development (Coombe, 2007), and some authors define them as “generative blueprints for test design” (Davidson and Lynch, 2002 in Coombe, 2007). The role of test specifications is also outlined in the Manual for Language Test Development and Examining, where test specifications are recognized to be of importance for both high-stakes and low-stakes assessments (Council of Europe, 2011). In the case of the former, test specifications are seen as an instrument for ensuring quality of a test and validity of inferences made on the basis of test results. Similarly, low-stakes assessments benefit from test specifications as well, especially in terms of ensuring that “all test forms have the same basis and that a test correctly relates to teaching syllabus ” (Council of Europe, 2011: 23). As suggested in the Manual, sample test specifications can be found in the works of Alderson, Clapham and Wall (1995), Bachman and Palmer (1996), and Davidson and Lynch (2002), but the author would like to propose using the CEFR as a basis for another test specifications model (see Milanović and Milanović, 2014).

The sample test specifications mentioned above will be discussed here as five widely used models which share some common characteristics, but it should be noted that they also

differ in various features. However, these models are not to be taken for the only possible and exclusive test specification models, although it can be argued that they provide test developers, test takers, and test users with useful pieces of information.

5.3.1.1.1 Alderson, Clapham and Wall (1995) Model

Although they are aware that some other authors use terms *test specifications* and *syllabus* interchangeably, Alderson *et al.* find differences between them. They argue that test specifications provide “the official statement about what the test tests and how it tests it” (1995: 9) and these can serve internal purposes of the examining body, which means that they are sometimes confidential, whereas the test syllabus, as a public document, contains information useful to teachers and test takers. Consequently, the former often contain valuable information for test and item writers, but they also provide test users, test takers and test validators with essential information for establishing test validity and usefulness (1995: 9). The stakeholders interested in test reliability and validity may have varying needs, so that Alderson *et al.* advocate using different forms of test specifications according to the type of audience that will be using them. Accordingly, they discuss test specifications developed for *test writers*, *test validators*, and *test users* respectively. Given the essential role of test and item writers in the process of test development, test specifications created to suit their needs are in the focus of our discussion here. As cited in Coombe (2007: 11-12), Alderson *et al.* include the following features into their model of test specifications intended for test and item writers:

- General statement of purpose
- Test battery (list of components and the time allowed for each)
- Test focus (description of the sub skills/knowledge areas to be tested)
- Source of texts (where appropriate text materials can be found)
- Test tasks (range of tasks to be used on the test)
- Item types (range of item types and number of items)
- Rubrics (form and content of instructions given to test takers).

Apart from test specifications developed for test writers, there is a recognized need for test specifications developed specifically for test validators and test users. Test validators’ role is

to provide arguments supporting validity of test results and inferences based on them, which means that they should be aware of the constructs the test intends to measure, as well as of the model of language ability these constructs are based on (Coombe, 2007). Test users, however, vary in their types of needs, although it is fairly easy to recognize several common types of users of test results: test takers, teachers (or educators), school/university officials, and employers. Alderson *et al.* suggest that test users should be made aware of what “the test measures, and what the test should be used for” (Alderson *et al.*, 1995: 20). Test specifications intended for test users are termed as “user specifications” and authors state that they should contain descriptions of a typical performance at each level, and also” a description of what a candidate can be expected to be able to do in the real life”. This is where the CEFR’s “can do” statements step in, because they are developed in such manner that they reflect a learner’s ability to use a target language (including grammar, vocabulary, and language functions) appropriately, while at the same time their performance can be linked to the corresponding levels on proficiency scales.

5.3.1.1.2 Bachman and Palmer (1996) Model

Bachman and Palmer argue that operationalization stage in test development consists of two interrelated activities(1996:171):

- 1) developing a *blueprint*, or the test specifications, and
- 2) developing test tasks and test task specifications.

In their model of test development, they distinguish between test specifications or *blueprint* that contains a detailed plan of the entire test and *test task specifications* (see 4.3.2 below), which is but a part of the blueprint. The blueprint can serve a number of purposes: (1) to permit the development of parallel forms of a test with the same characteristics, (2) to evaluate the work of test writers, (3) to evaluate the correspondence between the final product and the original intentions, and (4) to evaluate test (tasks) authenticity (Bachman and Palmer, 1996: 176-7).

The two-part specifications include the *structure* of a particular test, while the second part is what authors term as the *test task specifications*. According to Bachman and Palmer, a test blueprint normally includes the following (p.176):

- the number of parts/tasks
- the salience of parts/tasks
- the sequence of parts/tasks
- the relative importance of parts/tasks
- the number of tasks per part.

Once the blueprint has been finalized, actual tests can be put together, taking into consideration the principles of test usefulness: construct validity, interactiveness, reliability, practicality, authenticity. According to the authors of the model, test developers start with specifications of different test task types that they want to include in an actual test, and then they decide “how best to combine these in a test” (ibid.).

5.3.1.1.3 Davidson and Lynch Model (2002)

The third model we discuss here is developed by Davidson and Lynch (2002). As the authors point out, their model is somewhat similar to that of Bachman and Palmer, although some components of the two models are organized and labeled differently, with the significant differences referring to Bachman and Palmer’s explicitly stated time allotment, instructions and scoring method (Davidson and Lynch, 2002: 30). The model presented by Davidson and Lynch builds on the earlier one, developed by Popham (1978), consisting of the following five components:

- general description (a brief summary statement about what is being tested and measured)
- prompt attributes
- response attributes
- sample item
- specification supplement

Davidson and Lynch state that test specifications are aimed at creating tests which measure the same skill(s) as specified in this document, through a set of similar test tasks and items. The information contained in test specifications helps teachers, test administrators, test takers, test writers, and test users understand what is tested by the test and how results may be appropriately used (Davidson and Lynch, 2002).

5.3.1.1.4 Douglas's Model

The three models discussed above are not the only possible models of test specifications. Douglas, for example, says that test specifications should contain, at minimum, the following components:

- a description of the test content, including the organization of the test, a description of the number and type of test tasks, time allotment for each task, and specifications for each test task/item type,
- the criteria for correctness
- sample tasks/items (Douglas, 2000: 110-113).

As can be seen above, there are many possible ways of writing specifications that cover the essential elements identified by Douglas (Douglas, 2000 in Weigle, 2002: 83) depending on the purpose of assessment and intended audience for whom specifications are developed.

5.3.1.1.5 The CEFR Model

As outlined above, developing test specifications is not only recommendable but often a necessary and valuable step in developing language assessments. In this chapter we will explore the possibilities of using the CEFR in developing test/task specifications. It can be noticed that the three models of test specifications discussed above are very much in consensus as to what test specifications should include, although they use different terminology and ordering to list and describe test specification components. What interests us here is whether the CEFR and publications related to it can help test developers (or “constructors”) in the process of developing test specifications for a particular assessment purpose.

First of all, it should be noted that the CEFR was developed in order to meet a number of purposes:

- for the specification of test contents and examinations;
- for stating the criteria to determine the attainment of learning objectives;
- and
- for describing the levels of proficiency in existing tests and examinations for the purpose of their mutual comparisons across different systems of qualifications. (COE, 2001 in Milanović and Milanović, 2014)

The Chapter 4 of the Framework provides descriptions of language use and users, and more specifically, it focuses on communicative language activities in terms of spoken and

written interaction and production. For this reason, test developers need to adapt the CEFR to their own needs and the first step in this process is to specify the domain of language use and the purpose of their test (ESOL, 2011: 19). To this end, the CEFR can help test developers by drawing their attention to one (or more) of the following domains (see chapter 5.1.1 above): personal, public, occupational, and educational (COE, 2001:45). The users of the Framework are advised to select domains with respect to the needs of the learners who will have to operate in them, but it is to be noted that, depending on a situation in which language is used, more than one domain may be involved (COE, 2001: 45). When it comes to situations, they can be termed as *target language use* (TLU) situations where various language tasks can be identified, which is of much use in defining constructs which will be measured in language tests. Table 5 of the Framework provides examples of domains, including a number of variables that can be found within them: locations, institutions, persons, objects, events, operations, and texts. Communicative themes, tasks and purposes, communicative language activities and strategies are illustrated as well. However, the authors of the table state that this table is just an illustration of situations that may arise in each of the domains they identify, and therefore it “has no claims to be exhaustive or final” (see COE, 2001: 46, 48-49, and ESOL, 2011: 18). Consequently, test developers will have to work out the TLUs of their choice, and identify important characteristics they want to incorporate in their test specifications or test task specifications (Bachman and Palmer’s test task characteristics framework could also be of help in this process, 1996). Decisions regarding time allotment, instructions for responding, test rubrics and sample items and tasks have to be made by test developers, considering the purpose of assessment and the audience for which test specifications are developed. However, the Framework provides test developers with some hints in section 4.6 that deals with “texts” (page 93) and in section 7.3 related to tasks and their characteristics (page 157). These can be made use of together with “the growing “toolkit” designed to help designers exploit the CEFR” (ESOL, 2011: 19). This refers to an increasing number of publications related to utilizing the CEFR, including the *Manual for Language Test Development and Examining. For the Use with the CEFR* (COE/ALTE, 2011), *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual* (2009); the validated *Can Do statements* provided by the Association of Language Testers in Europe (ALTE); the publications and resources of the *English Profile Programme* (including the validated English Vocabulary

Profile wordlists, and the Can-Do statements for C levels of language proficiency- which are still the work in progress).

To sum up, it can be noted that the CEFR can provide valuable resources for test developers but it does not contain all the answers test developers may ask themselves in the process of developing a communicative language test (Milanović and Milanović, 2014).

5.3.2 Test task specifications

Palmer and Bachman argue that *a task* is the elemental unit of a language test, and for this reason test operationalization stage should focus on development of test tasks (1996: 171). Test tasks are developed with respect to target language use (TLU) task types in order to provide information on a test taker's ability to perform desired language functions in the real life. The starting point in test tasks development refers to identifying TLU task types which will provide a basis for the development of test tasks. The characteristics of test tasks should correspond to TLU task characteristics, and for this reason the latter should be identified and taken into consideration in the process of test development. Bachman and Palmer suggest that there are two strategies when it comes to writing actual test tasks (1996: 174):

- modify TLU tasks, or
- create original test tasks whose characteristics correspond to TLU tasks

The first approach, or strategy, that of modifying TLU tasks, can be taken when some characteristics of TLU tasks can be easily transferred to test tasks, but with certain modifications. For example, suppose the purpose of assessment requires a short speaking task for students enrolling in an undergraduate engineering course. It is relatively easy to identify a TLU task, such as giving an individual presentation on a course-related topic. However, due to the length of the preparation process and the TLU task in its entirety, it would be impractical to replicate all the characteristics of the TLU task. Instead, what test developers could do is to prepare a short prompt, based on which test takers could make an outline of the presentation and deliver it orally. The advantage of this approach is related to enhancing the authenticity of the assessment and its relevance to a TLU situation by “maintaining those characteristics of TLU task types that are considered to be distinctive” (p.176). Indeed, in communicative language testing, and in ESP language testing (which is communicative by definition, according to Douglas, 2000: 9),

authenticity is given a lot of significance, because of the test results which are used to make inferences about test takers' ability to use the language in the target language use situation.

In other situations, it may not be possible to identify TLU tasks which can be used as such, so in that case test developers need to consider their distinctive characteristics and then develop original test tasks sharing the same characteristics as the TLU tasks.

The TLU characteristics identified here are accompanied by the specific purpose and construct definition for each type of task which finds its way in a particular test, within a document known as *test task specifications* (Bachman and Palmer, 1996: 172). The authors argue that test task specifications need to include all of the following characteristics (not necessarily in the same order):

- 1) the purpose of the test task,
- 2) the definition of the construct to be measured (by a particular task),
- 3) the characteristics of the setting of the test task,
- 4) time allotment,
- 5) instructions for responding to the task,
- 6) characteristics of input, response, and relationship between input and response, and
- 7) scoring method.

5.3.3 Scoring method

Considering the fact that in any summative assessment scores are used to make inferences about test takers' language abilities and their language knowledge, and that these inferences are then further used for making certain decisions about test takers, the scoring method has to be well devised to suit the purpose of assessment. The method used to arrive at scores plays the most important part in the measurement process because of its role in securing the reliability of rating and validity of scores and inferences based on them (1996). To secure test score reliability, Bachman and Palmer suggest that test developers should follow three steps during the test development process:

- 1) define the construct theoretically,
- 2) define the construct operationally, and
- 3) establish the method for quantifying responses to test tasks (1996: 193).

Theoretical definition of the construct influences the type of score produced via the selected scoring method. There are three possible score types: a single composite score, the profile of scores, or a combination of the former and the latter. Decisions regarding the scoring method will be made during the operationalization stage, because it is at this stage that test task types are selected and developed.

In the operationalization stage, test developers have to make a number of different decisions. First, they need to consider test task types which will be based either on syllabus or a theoretical model, or both. Then they need to consider whether the tasks will cover the units listed in the syllabus, or they will be more related to TLU. Finally, they have to make decisions regarding the intended response, because it is the response to test task that determines the scoring method and the type of scores to be reported to test users.

Once the test has been developed and tasks and items have been included in the test, test developers have to address the issue of the most appropriate scoring method. It is generally accepted that some task types allow for more or less objective, while others allow for more subjective scoring. However, this distinction is not always black or white. Usually, it takes a great deal of pretesting and piloting of items to ensure that there is only one or a definite number of correct answers to tasks where the expected response is a selected answer. On the other hand, in some cases it is not possible to have a full control over the expected response, because test takers are prompted to respond by limited or extended production (as is the case with assessing writing and speaking) so that test developers must provide assessors with scoring scale(s) to ensure fair and objective rating process.

Scoring scales are mainly used to distinguish between different performances on test tasks, especially in those assessments which are prone to subjective rating, such as speaking and writing. It cannot be claimed that there is a universal terminology used to discuss the scales, as another term – rating rubrics - is used to denote more or less the same notion. However, a distinction can be made between these two, depending on their content, the intended use of the scales/rubrics, as well as on the intended audience. Regarding the content of scoring scales, it can be noted that they can be verbally or numerically described with the same purpose on mind – assess the performance and determine the scores which will “express *how well* the examinees can speak the language being tested” (Luoma, 2004: 59). Inevitably, the mere score (e.g. from 1,

being the lowest, to 5, being the highest score on a particular test) standing alone is not informative enough, although it can be otherwise expressed in verbal categories (e.g. *poor, fair, good, very good, and excellent*). To complement the meaning of the score, there are usually shorter or longer descriptions or statements developed so as to describe what characteristics of the performance the score refers to. In the case of speaking assessments, such descriptions, especially since they are ordered according to different levels - ascending or descending - are used to rate a performance and are referred to as rating scales or speaking scales (ibid. p. 59).

If the scales are used by raters, or examiners, to rate the test takers' performance, it seems reasonable to conclude that the term *rubrics* fits better than the *scales*; however, if the scales are to be used by test users, i.e. by people and institutions who will make certain decisions regarding test takers based on the test scores which is interpreted by the scales, or if the scales are to be used by test takers themselves to help them monitor their own progress; in this case it seems more appropriate to keep the term "scales". However, in this dissertation, the term "scales" will be used in all instances, because "rubric" will be used in Bachman's sense to talk about "characteristics that specify how test takers are expected to proceed in taking the test" (1990: 118), and include task instructions, time allotted for each task (and the whole test), and the organization of the test (test sections, and parts within the sections). This distinction between scales and rubrics will be of importance later in the dissertation because the research methodology makes use of Bachman and Palmer's *Test Task Characteristics Framework* (1996: 49-50), with some modifications to it made by Douglas (2000: 51-52) to analyze specific purpose language tasks and specific purpose test tasks.

The discussion above brings to light the important consideration in the process of scales development, and that is the audience for which the scales are written and developed. Some authors, like Luoma, distinguish among: examinees or test-takers, raters, and test administrators (2004). Bachman and Palmer, on the other hand, use the term "test users" instead of test administrators, to talk about teachers in educational systems, or potential employers outside university settings, as both the former and the latter "use" test results to make inferences about test takers' language ability for various purposes – placing students across levels according to their language ability, monitoring their progress, making hiring decisions in line with the job requirements and language needs in a particular company (Bachman and Palmer, 1996).

Speaking of using test results for making predictions about an individual's future performance in jobs that may require the use of a foreign language, McNamarra (1996) states that scores on language tests can inform two kinds of decisions – “(1) inferences about an individual's capability to perform future tasks or jobs that require the language use, and (2) inferences only about an individual's ability to use *language* in future tasks or jobs” (in Bachman and Palmer, 1996: 96). If the language test is to inform the decisions whether a candidate is suitable for a certain position in a company, test construct will inevitably have to contain elements pertaining to the characteristics, skills and topical knowledge necessary for completing the job-related tasks. On the other hand, if a test is supposed to inform decisions whether a candidate has the language ability to perform certain job-related tasks, the test construct will have to contain the considerations of individual characteristics that candidates need to possess in order to be selected for the position. When determining this, Bachman and Palmer suggest consulting subject matter specialists, e.g. an HR officer responsible for selection and recruitment in a particular company or industry, throughout the process of designing test tasks and developing the test (1996: 96). The consideration of audience for which the scale is created for will have certain implications on the complexity and wording of the descriptors within the scale. Luoma suggests re-writing scales for test-takers and test users in order to avoid technical terms and complex descriptions which are of use only to raters (2004: 83). Given that one of the intended purposes of this dissertation is to consider the implication of authentic test task formats on assessment development and its future influence on test takers' ability to use the language within labor markets, the term *test user* will be adopted together with those of a *test-taker* and a *rater*, when discussing speaking scales used for scoring and interpreting test takers' performance in oral assessments.

5.3.3.1 Rating scales

Scoring test takers' performance in oral assessment can be problematic for several reasons. First of all, some of the most interesting items to score call for “the most complex kinds of subjective scoring” (Cohen, 1994: 87). In order to avoid subjectivity, test developers create scales simultaneously with developing test tasks for speaking assessments. In most assessments there are different scales intended to target different audiences: raters, test takers, and test users. These scales differ in the quantity of information they offer to the respective audiences, the terminology used to describe the test takers' performance, and in “the focus in terms of *what* the

examinees can do and *how well* they can do it” (Council of Europe 2001, in Luoma 2004:60). In the same vein, Alderson (1991) makes a functional distinction between three types of proficiency scales: *user-oriented* (they report typical behaviors of learners at any given level focusing on **what a learner can do**), *assessor-oriented* (they guide the rating process, and although they are often negatively worded, descriptions of reference levels can follow the example provided in Table 3 of the Framework and employ positive wording with necessary limitations in establishing **how well a learner performs**) (COE, 2001: 28-29), and *constructor-oriented* (they inform the process of test development at appropriate levels of proficiency by providing statements expressed in terms of specific communication tasks the learner is to perform in a test, demonstrating what they **can do**). A problem may occur if proficiency scales designed for one function is used for another (2001: 37), for example if raters use user-oriented scales to evaluate performance (in Milanović and Milanović, 2014). The most comprehensive scales are used by raters, and it is these scales that will be of primary concern in the rest of the discussion.

Raters use scales to assess how well a candidate completes a given task, in order to reduce any possible effects of subjective marking. To do this, they adhere to scales containing different levels and descriptors explaining what each level should mean. In other words, they explain what kind of performance can be expected of test takers at each level. Recognizing the performance and matching it to the corresponding descriptor in the scale, or a “statement of the kind of behavior that each point on the scale refers to (Alderson et al., 1995: 107), is a primary consideration in the rating process. Based on the number of levels and categories that they cover as well as on the judgments that they help to be formed, rating scales can be divided into two broad categories – **holistic** and **analytic**. The difference between the two is not only related to the number of levels and categories, but it is reflected in the score derived from the analysis of the performance as well. *Holistic assessment* is all about making a “global synthetic judgment” by using holistic scales that are suitable for rating the overall effectiveness of test takers’ performance (Council of Europe, 2001: 190). Some raters opt for them because they combine descriptors reflecting a mix of abilities within a level, they are faster to use because there are fewer criteria and are considerably easier to apply because there is not much material that raters should remember while assessing the performance (see sample holistic scale, Table 5.3 below). They give a single score, which is useful for many purposes. However, it can be argued that they are more useful to test users and test-takers who use them to analyze the test-takers’ overall

performance than to raters who endeavor to identify individual strengths and weaknesses in a performance. As North states (1994), they are very much dependent on quantifiers such as *a few*, *some*, *many*, as well as quality words such as *sufficient*, *relevant*, which bear different meanings to different raters (see sample holistic scale, Table XXXX below); or even different meanings to the same raters on various occasions, thus potentially affecting intra- and inter-rater reliability, as well as jeopardizing validity of inferences based on the scores. Holistic scales are often referred to as *global* scales (Douglas, 2000:71), because they offer a more general view of the demonstrated ability, and *impression* scales, in cases when a decision has to be made rapidly (Alderson et al., 1995: 108). Perhaps these two terms – global and holistic – best demonstrate potential uses of holistic scales in testing situations when it is necessary to make fast decisions about overall performance, without lengthy standardization sessions for rater training. On the other hand, the convenience for use comes with a price because holistic scales fail to reflect nuances in performance which would offer a more comprehensive picture of test takers’ ability.

Table. 5.3: A Sample Holistic Scale (From UCLES International Examinations in English as a Foreign Language General Handbook, 1987 in Alderson et al., 1995: 107)

18-20	Excellent	Natural English with minimal errors and complete realization of the task set.
16-17	Very Good	More than a collection of simple sentences, with good vocabulary and structures. Some non-basic errors.
12-15	Good	Simple but accurate realization of the task set with sufficient naturalness of English and not many errors.
8-11	Pass	Reasonably correct but awkward and non-communicating OR fair and natural treatment of subject, with some serious errors.
5-7	Weak	Original vocabulary and grammar both inadequate to the subject.
0-4	Very poor	Incoherent. Errors show lack of basic knowledge of English.

Holistic scales are of little use when performance has to be analyzed with regards to various components (e.g. fluency, pronunciation, accuracy, vocabulary use, etc.) or to different aspects separately. In this case, raters opt for *analytic* scales, which help them place the test

takers' performance at a particular level, or a band on the scale. Two important considerations that test developers face when developing rating scales refer to the number of levels and criteria to be included in the scales. Given that consistency in rating performance is essential for securing reliability and validity of test results, it is important to note that raters cannot distinguish consistently among too many criteria in a scale. Test-specific scales usually have 4 to 6 levels, which are labeled by numbers, percentages, or level markings (e.g. A1, A2, B1, etc.). Criteria, on the other hand, contain scale descriptors explaining the kind of a performance that can be expected of test takers at each level, thus giving meaning to different levels on the scale. According to Council of Europe, 4-5 categories cause a cognitive overload for raters, while 7 categories should be regarded as an upper limit above which raters can no longer distinguish among various aspects of performance (2001: 193). As a consequence, a large number of criteria for scoring might not yield consistent ratings, which will affect the reliability of scoring and validity of inferences based on the test scores. Once the number of levels and criteria has been decided, raters can focus on deciding "how far up the scale test takers can go", meaning that there is a vertical emphasis in using the scale (ibid. p. 189). As opposed to holistic scales that derive one, composite score, analytic scales offer a profile of scores.

Table 5.4: The test of Spoken English band descriptors for Overall features (ETS, 2001b:30 in Luoma, 2004: 70)

0	<p>Communication almost always effective: task performed very competently.</p> <p>Speaker volunteers information freely, with little or no effort, and may go beyond the task by using additional appropriate functions.</p> <ul style="list-style-type: none"> - Native-like repair strategies - Sophisticated expressions - Very strong content - Almost no listener effort required
0	<p>Communication generally effective: task performed competently.</p> <p>Speaker volunteers information, sometimes with effort; usually does not run out of time.</p> <ul style="list-style-type: none"> - Linguistic weaknesses may necessitate some repair strategies that may be slightly distracting. - Expressions sometimes awkward - Generally strong content - Little listener effort required
0	<p>Communication somewhat effective: task performed somewhat competently.</p> <p>Speaker responds with effort; sometimes provides limited speech sample and sometimes runs out of time.</p>

	<ul style="list-style-type: none"> - Sometimes excessive, distracting and ineffective repair strategies used to compensate for linguistic weaknesses (e.g. vocabulary and/or grammar) - Adequate content - Some listener effort required
0	<p>Communication generally not effective: task generally performed poorly.</p> <p>Speaker responds with much effort; provides limited speech sample and often runs out of time.</p> <ul style="list-style-type: none"> - Repair strategies excessive, very distracting and ineffective - Much listener effort required - Difficult to tell if task is fully performed because of linguistic weaknesses, but function can be identified
0	<p>No effective communication: no evidence of ability to perform task.</p> <p>Extreme speaker effort is evident; speaker may repeat prompt, give up on task, or be silent.</p> <ul style="list-style-type: none"> - Attempts to perform task end in failure - Only isolated words or phrases intelligible, even with much listener effort - Function cannot be identified

Practically, this means that test raters use a checklist to map test takers' performance (see an example of analytic score, Table XXX below), whereas test takers can get different scores for different criteria, and if the performance has to be expressed in a single score, the way that criteria and tasks are weighed will determine on the strategy applied in obtaining a single score (expressed with an illustrative grade, letter, number, or a percentage).

Depending on the purpose of the assessment and the TLU, rating scales can be based on a theoretical model of language acquisition, in which case we talk about *theory-derived* scales. Such scales describe “degrees of language ability without reference to specific situations”, and are mainly based on the model of communicative competence, such is Bachman and Palmer's (1996) Communicative Language Ability Model (Luoma, 2004: 67). When scales are developed to help assessing response on a particular task developed in accordance with the corresponding TLU task, they refer to a specific situation and describe linguistic and non-linguistic performance on the task. In such case, we talk about *behavioral scales* that raters use to assess test takers' performance. Furthermore, behavioral scales can be useful for describing various tasks that test takers can be expected to demonstrate at different levels specifying “the degree of skill with which they can handle them” (ibid. p. 67).

5.3.3.2 *The CEFR Scales*

The CEFR is concerned with language assessment in terms of providing solid basis for ensuring validity, reliability, and feasibility of assessments, so its authors suggest it be used in the following three ways:

- for the specification of test contents and examinations;
- for stating the criteria to determine the attainment of learning objectives; and
- for describing the levels of proficiency in existing tests and examinations for the purpose of their mutual comparisons across different systems of qualifications. (COE, 2000 in Milanović and Milanović, 2014)

In other words, the Framework may help test developers, administrators, secondary and higher education officials to determine what is assessed, how performance is interpreted, and how comparisons can be made. However, there is some criticism of the Framework regarding its application in test specifications development. First of all, the critics claim that the CEFR is not a framework but a model of language proficiency, which is too abstract to enable test developers to write test specifications that will mirror the Framework (Weir, 2005, Fulcher, 2004). Fulcher argues that “true frameworks need to mediate between the abstract and the context of a particular test” with the purpose of operationalizing the components of a model which are in line with a specific purpose of a test, and as such the framework enables test developers to produce test specifications (Fulcher, 2004: 259).

The Framework scales are argued to be helpful to providing performance descriptors that will find their place in scales used to rate performance. However, care must be taken to distinguish between descriptors of communicative activities and descriptors of aspects of proficiency related to particular competencies. The former can be useful for reporting results to test users (employers, university officials and administrators, etc.), whereas the descriptors of aspects of proficiency related to particular skills and competences may be used for specifying criteria for performance assessment. The latter can be done in three ways:

- descriptors can be presented as a scale in the form of a holistic paragraph per any given level,
- descriptors can be presented as a checklist where descriptors are grouped under categories, and
- descriptors can be presented as a checklist of selected categories, which makes it possible to give a diagnostic profile. The checklist of sub-scales can take the form of proficiency scale, where relevant levels are defined for certain categories, and it can take the form of an examination rating scale, where descriptors are defined for each relevant category (ibid.).

The scales of descriptors provided by the Framework can be of use to the process of language assessment provided there is an accurate identification of the purpose the scale is to serve (COE, 2001: 6). The Common European Framework for Reference (CEFR) is often used as a tool for educators and assessors because, despite being language-neutral, it contains a number of scales whose descriptors cover a number of situations where language is used for reception, interaction and production. In the context of language assessment, the CEFR descriptors can be used to create test-specific criteria, since they cover various aspects of performance on a task, including a comprehensive list of descriptors dealing with linguistic features of learners' output. Perhaps, this is best exemplified in Table 5.5 below, which contains a set of descriptors on an analytic, behavioral rating scale, which can be used to map spoken language use in an assessment.

Table 5.5 Qualitative aspects of spoken language use (Council of Europe, 2001: 28-29)

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical Control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turn taking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing Controlled use of organisational patterns, connectors and cohesive devices.

B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical Control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he / she needs to, though he /she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circum-locutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and", "but" and "because".
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited Control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then".

This scale is not written for any specific language or purpose, so if it is to be used in a specific assessment, its descriptors have to be modified so as to suit the purpose of assessment and the tasks which it will be used to help assessing. Furthermore, as Luoma observes, test developers, who opt for this scale, would have to decide how to derive a score. For example, they can derive five analytic scores, an overall score considering all five criteria, or both scores (2004: 71). It should also be noted that although the CEFR offers numerous descriptors that can be used in scale writing, some authors have found them to be vague and inconsistent. Alderson *et al.* found similar descriptors occurring at different levels, different verbs describing apparently the same cognitive process, etc. (Alderson *et al.*, 2004 in Weir, 2005: 16-17).

6 Research methodology

This chapter offers a brief overview of research goals and the instruments applied to achieve the goals. The subsequent subchapters outline the main research questions, the author's hypotheses and expected results, and, finally, research instruments employed to find answers to the research questions.

6.1 Research questions

As outlined in Chapter 1, the study endeavors to find the answers to the following research questions:

- 1) Can target language use situation tasks be used as a model for authentic classroom test tasks?
- 2) Do authentic forms of assessment exert a positive influence on students' progress?
- 3) Should background knowledge be tested in specific purpose speaking assessments?
- 4) Do authentic forms of assessment exert a positive influence on students' awareness of their own progress?
- 5) Do business students possess the language skills matching the needs of the labor market?

To provide answers to the research questions listed above, the author conducted a research divided into 2 phases:

Phase 1 – collecting data in collaboration with 25 subject specialist informants representing the real life domain (labor market at the territory of the Municipality of Kragujevac); and

Phase 2 – collecting data in the domain of higher education, on the sample of 150 business students enrolled in the Faculty of Economics (modules: Management , Accounting and Business Finance, and Marketing), University of Kragujevac.

The data collected during the two phases of the research were analyzed and used to test and validate the hypotheses presented in the following chapter (see Chapter 6.2 below).

6.2 Hypotheses and expected results

6.2.1 Hypotheses

The research conducted for the purposes of this doctoral thesis aims at investigating spoken English language skills assessed through formative and summative test methods, by the means of authentic input material and test tasks. The test tasks used in the research come as a

product of a thorough analysis of target language use situations in which language users complete various real life language tasks (Bachman and Palmer, 1996). In this way, the author will investigate authenticity of test tasks that are created based on the TLU situation analysis, as well as the effect that authentic speaking tasks have on students' progress. Bearing in mind that class assessment within any particular curriculum has two purposes – to check both student progress and attainment of learning objectives, and to ensure that future employers' expectations are met – the research aims at determining the extent to which authentic test tasks may have a formative role in facilitating students' progress.

Based on the theoretical framework presented in the first part of the dissertation, an empirical research will be conducted with the purpose of testing and validating the following:

H1: The examinees who have been thoroughly trained to apply evaluation criteria demonstrate a better overall performance in the final oral exam in comparison to the examinees who have not been thoroughly trained on applying analytic and holistic scoring criteria in assessing their own and the performance of their peers.

H2: Performing on a task requiring that test takers should possess background knowledge related to the field of *Marketing*, the Control group demonstrates very similar results to the more successful of the two experimental groups.

H3: End of semester survey results indicate that more than two thirds of the examinees demonstrate positive perceptions of authentic tasks, as well as of the system of evaluation and self-evaluation that they have been exposed to.

H4: End of semester self-evaluation questionnaire results indicate that at least 70% of the Control group's responses provided to estimate their target skills match the responses provided at the beginning of the semester.

H5: End-of-semester self-evaluation results indicate that at least half of the sample in the Experimental groups achieved progress by one CEFR level, as corroborated by the Second placement test results.

H6:H6: The highest agreement in responses to the “Can-do” survey is the one between subject specialist informants and Group 1 subjects.

6.2.2 Expected results

H1: It is expected that this research will prove that being familiar with criteria for correctness has a formative impact on learning, facilitating future performance inside and outside the classroom settings. When it comes to the target language use situations, this hypothesis is relevant in the sense that at every workplace there is a set of implicit and explicit criteria that employees follow in order to complete tasks.

H2: In many assessments, the influence of background knowledge may contaminate the score unless the background knowledge is a part of the construct as is the case with many specific purpose language assessments. To this end, proving the validity of H2 intends to show that background knowledge may play an important role in carrying out certain tasks, especially the tasks in which communicative goals require that speakers demonstrate more than just the ability to communicate in a foreign language. This particular feature is of importance in teaching and assessing languages for specific purposes since the vocabulary taught in these settings is always field-specific and requires that learners should use it bearing in mind the specific context it is associated to.

H3: By checking the validity of H3, the research will provide insight into student perceptions of authentic forms of assessment, as well as of authentic forms of evaluation. Given the relationship between learner's perceptions of the assessment and their motivation to achieve deep learning (Trigwell and Prosser, 1991 in Struyven et al., 2004:26), these findings will reveal whether students recognize authentic forms of assessment as valuable and important to their learning process.

H4: It is an assumption on behalf of the author that if examinees are not exposed to authentic tasks as well as to authentic forms of assessment and evaluation, they will less likely be aware of their own progress, as well as of their current language ability. Consequently, they will less likely be efficient in keeping track of and recognizing their progress by using the self-evaluation checklist containing the CEFR-aligned descriptors of spoken interaction and production. The author assumes that students in the Control group will demonstrate the lack of self-awareness when it comes to their own progress, documented by setting the same targets at the end of the semester. It is also assumed that students who receive a detailed feedback and learn how to interpret evaluation criteria raise self-awareness and the sense of what area of their

performance needs improvement. In this way, the research will prove the relationship between authentic forms of evaluation and the ability to self-monitor own progress in learning a foreign language.

H5: By checking the validity of H5, the author of the thesis wants to point out that authentic assessment forms exerted a positive influence on the examinees' performance, including their capability of estimating their own performance, underlining a positive, formative nature of authentic test tasks and forms of evaluation.

H6: The author assumes that Group 1 respondents, exposed to authentic test tasks and trained on assessing performance by using a detailed, analytic rating scale, possess the skills similar to those required in work settings. Their performance on the tasks requires collaboration and peer-coaching, emulating the characteristics of work settings.

If the author's assumption that authentic test tasks and test performance evaluation methods correlate with target language use tasks and methods of evaluation is proved to be true, the conclusion to be drawn is that such forms of assessment play a formative role bringing students' language skills closer to the requirements of the labor market. Employers have certain expectations of the language skills their prospective employees should possess before they join the company, so it is university where these skills need to be developed.

The research results should point out strengths and weaknesses of the English language 2 course syllabus at the Faculty of Economics, University of Kragujevac. If the research results show that authentic test tasks and evaluation exert a positive influence on student learning and that they stimulate learning, the syllabus will undergo certain changes so as to be more relevant to target language use situations.

6.3 Data collection and instruments

To test the hypotheses the research employs data collected from two groups of participants in two phases of data collection:

- Phase 1: subject specialist informants (representing the real life domain), and
- Phase 2: business students (representing the educational domain).

In line with the research rationale, in Phase 1, the author will rely on the help of *subject specialist informants* (HR officers, managers, PR managers, etc.) employing business graduates in privately-owned companies at the territory of the Municipality of Kragujevac) in analyzing the real life language use in target language use situations, by applying *context-based research* technique, and by following the principles of *grounded ethnography* qualitative research. Having analyzed the context, the author will apply the *Task characteristics framework*, developed by Bachman and Palmer (1996: 49-50) and further modified by Douglas so as to suit the specific purpose language assessment (Douglas, 2000:51-52), in order to analyze speaking tasks within the TLU situations. The findings obtained in this manner are then used to develop speaking (test) tasks and tasks specifications for the use within the educational domain with another group of participants in the research – business students. It is assumed that speaking test tasks developed in this way share the characteristics of the target language use speaking tasks, including the characteristics of situational and interactional authenticity (Douglas, 2000: 14).

The speaking test tasks, developed during the first stage of the research, are administered in the second stage (Phase 2). Student participants, who signed the consent forms and agreed to participate in the research, will be subjected to formative and summative language assessment procedures by being exposed to the following authentic assessment forms: speaking test tasks that share situational and interactional authenticity with TLU tasks, self- and peer assessment, feedback, and self-monitoring by the application of Can-do self-evaluation checklists. To validate the research hypotheses, results obtained by the assessment procedures will be corroborated at the end of the semester and statistically analyzed by the administration of the following statistical instruments:

- The Kolmogorov-Smirnov Test,
- The Mann Whitney Test,
- The Sign Test,
- The Kruskal-Wallis Test,
- The Pearson Chi-Square Test,
- The Kappa Test,
- The Shapiro-Wilk Test,
- The Hosmer and Lemeshow Test,

- The Nagelkerke R-Square Value, and

The results of the statistical analyses and discussion are presented in Chapter 9 below.

7 Phase 1

7.1 From target language use to test tasks

The approach adopted in the research follows Douglas's guidelines for identifying and specifying language tasks in a specific purpose TLU domain by "investigating and describing the target language use situations that form the basis for identifying specific purpose test tasks" (Douglas, 2000:92). The process of developing specific purpose speaking test tasks is a challenging one in that it requires that the following be considered:

- identifying the target language use domain,
- identifying target language use situations within the domain,
- identifying possible speaking tasks within situations,
- describing task characteristics and translating them into test task characteristics,
- developing test/ test tasks whose characteristics correspond to TLU speaking tasks.

7.1.1 Identifying the target language use domain

The starting point in the LSP testing refers to the analysis of the target language use domain, situations that occur within the domain, and specific language tasks that language users perform within the TLU situations. The target language use domain relevant to this research is what Bachman and Palmer refer to as the real life domain (1996); it can also be referred to as the occupational domain according to the Council of Europe's classification (2000). This domain can be narrowed down to the domain of business, referring to the business operations performed by small and medium enterprises on the territory of the Municipality of Kragujevac in Serbia. The research focuses on the companies which perform business both locally and internationally, making it a requirement for their employees to be able to use spoken (and written) English language for business communication on a daily basis. In addition to this, the research pinpoints the target group of business graduates (seeking jobs within the TLU domain) whose spoken English language skills are the subject of observation and assessment.

7.1.2 Identifying target language use situations and special purpose speaking tasks

Given the importance that context and target language use situations have in developing test tasks in the light of language for specific purpose language tests, we will consider a couple of techniques that are used in analyzing the context and TLU tasks with the purpose of providing solid foundations for the development of specific purpose test tasks. The following techniques will be discussed: *grounded ethnography*, *context-based research*, and *subject specialist informant procedures*.

Douglas refers to the aforementioned techniques as to the techniques that LSP test developers need to consider when they analyze target language use domains, aware that domains and the corresponding situations have an immense number of variables that are difficult to predict and control in a specific purpose language test. Another problem inherent to the LSP testing refers to test developers often being unfamiliar with the specific purpose field to which test scores are supposed to generalize. To overcome this problem, test developers seek help from an expert in the field to clarify the specificities of the TLU situation and the type of communication that takes place in it. This help is crucial to the process of test task development, if the test is supposed to claim any relevance to the target language use situation. The following techniques will be discussed in the subsequent chapters: grounded ethnography, context-based research, and subject specialist informant procedures.

7.1.2.1 Grounded ethnography

Ethnography, as an approach to studying human behavior, which appeared in the late 1960s, influenced the work of Frankel and Bechman who utilized the ethnographic research techniques to develop a technique for studying human behavior in context, i.e. in a particular situation. They define the technique as follows:

a means for the researcher to understand an event by studying both its natural occurrence and the accounts and descriptions of it provided by co-participants.

(Frankel and Bechman, 1982 in Douglas, 2000: 93)

Ethnographic research, being a qualitative, process-oriented, research technique, deals with detailed descriptions of a context- setting, time, participants, purpose, etc. Its purpose is to provide a detailed account of the context, human behavior and interactions, including the interpretations of the language and behavior resulting from and in the context. Frankel and

Bechman built on the idea of ethnography by adding a videotaping element to it, naming the technique *grounded ethnography* (ibid.). Ideally, observers are, at the same time, participants in the context that they are trying to analyze. However, since this may cause awkward and artificial behavior of other participants, the authors suggest videotaping situations and then observing them upon their completion, in collaboration with other experts. They argue that there are several advantages to this procedure:

- by watching the recordings, the participants can make direct comments, without having to recall the situation in question (as would be the case if accounts were memory-based);
- by using “hidden or inconspicuous cameras” the researchers avoid intrusion characteristic of a situation where they are taking the role of co-participants;
- expert commentaries (from linguists, ethnographers, field specialists, etc.) allow for being audio-recorded, transcribed and inserted into the recording transcripts, facilitating further analyses;
- indigenous assessment criteria [in Douglas’s sense, 2000] can be brought to the researchers’ attention;
- differing viewpoints can add to having a better insight into the TLU situation and its specificities. (ibid.)

The approach and its original design are supposed to help test developers in analyzing varying aspects of TLU contexts so that they can develop test tasks sharing similar characteristics to those of the situation in the TLU context. However, we must point out that there are certain limitations to this approach. First, the idea of videotaping participants in a particular real life situation, even with their consent, is somewhat problematic in the light of 21st century data protection laws and various confidentiality-related regulations adopted by companies, agencies, and other legal entities. Second, the quality of practicality poses numerous constraints on test development teams, limiting their resources in terms of personnel, budget, and time. The author of the thesis admits that this problem is somewhat alleviated in professional testing associations which can afford to allocate resources for hiring field-specific experts according to their particular needs. Third, when it comes to analyzing indigenous assessment criteria (see Chapter 5.2.1.5), participants in real life situations often do not possess adequate abilities of explicitly stating the assessment criteria they employed in a particular situation. Naturally, in particular situations, participants make internal judgments and evaluate other people’s words and actions, but they may not be particularly “useful” in reporting the criteria they had applied in making the assessments. Another problem to indigenous assessment criteria refers to their being “highly contextualized and task-specific”, making it difficult for test

developers to transfer them successfully to a language testing context (Jacoby and McNamara 1999, in Douglas, 2000).

7.1.2.2 *Context-based research*

An important aspect to analyzing TLU domains resulted from Douglas and Selinker's (1994) work on what they call "*context-based*" research. Building on the ideas of grounded ethnography and subject specialist informant techniques, they provided guidelines for context-based research, which they define as a "study of second language acquisition and use in real life contexts" (in Douglas, 2000:95). This technique takes into consideration two kinds of data: primary and secondary. *Primary data* result from empirical study providing researchers with "the interlanguage talk or writing" they wish to study. According to the principles of grounded ethnography approach, participants in the observed context, giving comments on the primary data, provide *secondary data*. Douglas and Selinker add on this idea, by differentiating between two sources of secondary data commentaries - the one coming from the very participants in a situation, and the other coming from various experts engaged in the process of data analysis (ibid.). The idea of a context-based research pinpoints the dynamic nature of context, to whose development participants contribute by their input. This input, created by the interlocutors' respective turns, or by the speaker and audience's respective characteristics, is not always easy to interpret by the participants in a particular situation. For this reason, test developers benefit from expert interpretation of the communication that takes place in a target language use situation. The technique involving help from subject specialist informants was developed to provide the required expertise.

7.1.2.3 *Subject specialist informant procedures*

One of the greatest challenges that LSP test developers face is the ignorance of the situation for which they are developing a specific purpose test. They may not feel certain what kind of data to focus on, what kind of performance is expected, or what aspects of data and performance the professionals in a particular field appreciate. To bridge this gap, Selinker (1979) argues that test developers should seek help from *subject specialist informants* to help them "understand input data in LSP disciplines with which the test developers have little or no expertise" (in Douglas, 2000:97). What constitutes a specific purpose situation, according to

Douglas, is not only special purpose terminology and special content, but also a context “created by the professionals who Control the content and language in purposive interaction” (p.98). Selinker proposes that subject specialist informants have a better insight of the language use in a particular field than language testers who have little or no experience with the TLU situation in question. In addition to this, Elder suggests that subject specialist informants, although they need not have any background in linguistics, are actually quite reliable assessors of specific purpose language ability because they focus on the achievement of communicative goals rather than on the language itself (in Douglas, 2000:99). As such, Douglas argues that they should be involved in the testing project from the very beginning (ibid.).

Regardless of the apparent advantages to utilizing the subject specialist informant approach, it should be noted that not every professional in a particular field is suitable for the role of an informant. As LSP developers intend to develop language tests, they need to rely on the informant’s judgment of linguistic performance as well as the use of technical language. For this reason, Douglas proposes collaborating with subject specialist informants who are: a) “sensitive to technical language, and b) tolerant on linguistically oriented questions (ibid.). To mitigate the potential problems arising between a test developer and informant, Huckin and Olsen (1984) suggest that they should reach common ground by the informant giving a “top-down understanding of the purpose of the LSP text or interaction and its main content”. Once they have reached the common ground, they may continue by working on the “lower level, bottom-up rhetorical and grammatical aspects” of data (in Douglas, 2000: 99-100).

Moving from theory to practice in this research, the author has applied Douglas’s theoretical framework for investigating target language use situations in the following manner:

- identify potential labor market representatives (HR managers, PR managers, managers) who could inform the research,
- assign them the roles of subject professionals and subject specialist informants,
- get help with identifying context – SP target language use situations and tasks,
- identify preferable target language use speaking skills,
- analyze target language use speaking tasks,
- develop speaking test tasks that will correspond to TLU tasks,
- apply speaking tasks in the educational domain, and

- test hypotheses by qualitative and quantitative research methods.

7.2 Participants in Phase 1

During the Phase 1 of data collection, the author contacted 42 companies registered at the territory of the Municipality of Kragujevac, requesting help with the research. This part of the research follows the guidelines of grounded-ethnography technique, including the principles of context-based approach, and the assistance of subject-specialist informants (Douglas, 2000). To be considered suitable for data collection, the contributors needed to meet certain requirements related to the type of a company they work for, and the professional profile of a delegate who contributes data on behalf of the company.

The companies were identified based on the following selection criteria:

- the selected company that hires recent business graduates majoring in *Marketing, Management*, or *Accounting and Business Finance*,
- the size of the company is in the range between 15 and 250 employees,
- it performs business operations in collaboration with foreign partners (import-export, franchising, authorized dealership, etc.),
- the selected company requires the use of spoken English language in addition to Serbian as the language of business communication.

The contacted companies were required to delegate a representative who could help the research by providing the following data (used anonymously in the research, based on the Data contribution consent form, Appendices F and G):

- 1) a brief description of target language use situations, taking place in their respective companies, where English is used as a medium of communication (by the means of closed-ended/open-ended questionnaire during the interview with the researcher); and
- 2) an indication of the desired level of proficiency in English, in terms of speaking skills, by responding to a closed-ended “can-do“ checklist provided by the researcher.

There was an additional requirement concerning the delegation of an appropriate company representative. A designated delegate was required to assume the dual role of:

- a professional who has a hands-on experience of the context;

- a subject-specialist informant who is able to provide a top-down perspective of the contexts in which employees use spoken English for business communication (by responding to the Context-based questionnaire), as well as a bottom-up perspective referring to the oral English skills relevant to attaining communicative goals (by responding to “Can-do” questionnaire).

The process of data collection involved interviewing the delegated representatives live or via Skype, asking them to provide answers to multiple choice questions, and short descriptions in response to open-ended questions, which were then recorded and analyzed by the researcher. Data collection process lasted from January 2015 until March 2016, involving 42 companies that met the requirements for the participation in the research. An additional requirement posed by the researcher was that responses to both questionnaires should be fully completed to be considered for data analysis. After the initial contact and an agreement to sign the consent form, 8 company representatives failed to set an appointment for the interview or respond to the questionnaires in writing. The remaining thirty-four respondents provided assistance by responding to the first questionnaire over an interview with the researcher who recorded the responses (see the Context-based questionnaire below, Table 7.1). In the end of Phase 1, the total of twenty-five interviewees responded to both the Context-based and “Can do” questionnaire (Appendix Q, parts A and B). As per the research requirements, only the full responses were eligible for further analysis (25 in total).

7.3 Research instruments

In Phase 1, the author collected data in collaboration with the representatives of the labor market who assumed the role of subject specialist informants feeding the research with data on language tasks that take place within target language use situations in the real life domain. There were 25 informants who provided complete answers to two questionnaires:

- the context-based research questionnaire, and
- the “Can-do” questionnaire.

The purpose of the context-based questionnaire is to analyze the contexts in which speaking tasks occur in the real life domain and use the data resulting from the analysis to emulate task characteristics in the process of test task development so that test tasks correspond to the real life tasks. In this way, test takers’ interaction with the task will share the same

characteristics with language users' reaction to the real life task. Both procedures are intended to increase the quality of authenticity.

The 'Can-do' questionnaire is administered both to labor market representatives and students (in Phase 2 of data collection). The labor market respondents filled out the questionnaire (in the form of an evaluation checklist) identifying the spoken English language skills that their prospective employees should possess (Appendix Q, parts A and B, and Appendix R with parts A and B, as a "key" to Appendix Q). The descriptors in the checklists, based on the Council of Europe's *Common European Framework of Reference for Languages*, are modified so as to suit the specific purpose language assessment. The same questionnaire was administered in the educational setting, in Phase 2, with descriptors shuffled in the same manner and for the same reason (the author will discuss data collection in Phase 2 below). It should be noted that the author intentionally excluded C1 and C2 descriptors since they are rare in the educational setting and very few respondents in Phase 2 demonstrated the ability at a level higher than B2.

Students responding to the "Can-do questionnaires did so with the intention to monitor their own progress. It should be noted that the labor market informant's responses to the "Can-do" questionnaires were collected during Phase 1, after which they were collated and statistically analyzed together with student respondents' data in Phase 2. The Mann-Whitney test was used to statistically analyze data for the purpose of testing and validating H4 (see Chapter 6.2.1).

The role of the subject specialist informants is to familiarize the researcher with the TLU context and help him obtain two important deliverables:

Deliverable 1 - TLU speaking task characteristics,

Deliverable 2 - the desired CEFR level for spoken production/interaction in English.

7.3.1 Description of a target language use situation

Aware that the task of providing a clear and coherent description of as many as possible situations in which English language is used in company settings is not be an easy task for the respondents to the survey, the researcher developed a Context-based questionnaire, combining open-ended (MCQ) and closed-ended questions allowing for more freedom of description (Table 7.1 below). The responses to the Context-based questionnaire provide valuable input for Douglas's TLU Task Characteristics Framework, which will be used as a basis for the

development of special purpose test tasks sharing the same characteristics as the TLU tasks. Considering that the research aims at investigating the extent to which the current test tasks are authentic in comparison to language use tasks outside testing and educational settings, this phase of the research aims at providing enough material for the development of test tasks which will elicit authentic responses (interactional authenticity) in authentic settings (situational authenticity). Bearing in mind that target language use situations differ even within a single business setting, the researcher conducted the survey disillusioned that all respondents would provide identical answers. Instead, the survey was conducted with the assumption that it would yield the most common language task characteristics. The survey results are collected, compared, and analyzed for the following purposes:

- identify and define situations in which various settings share the common ground (similar problems/situations when speakers use oral English skills as a medium of communication),
- identify speaking task formats that are in the same or similar format across settings,
- analyze tasks using the Task characteristics framework.

7.3.1.1 Context-based survey

The respondents to the survey (25 in total) provided their answers orally to the researcher who recorded them (in English) in writing, for the purpose of further analysis. The questionnaire consists of two parts: General context and Business presentations. General context questionnaire (see Appendix H) provides a general idea of the business setting that the respondent comes from: the type and size of their company; questions related to prospective employees with educational background in business; questions related to the use of spoken English for the purpose of business communication; the type of oral performance (production or interaction); and the relative importance of speaking tasks delivered in English. Item number 7 is a ranking question with seven options, prompting the respondent to rank them in the order of importance (1 being the most important, 7 being the least important). Given that the survey took the form of an interview, the researcher explained the prompt and the ranking system to the respondents.

Table 7.1. Context-based questionnaire: Part 1 - General context

1	Your company performs business_____.	a) locally	b) internationally	c) both a and b
2	Your employees are required to use spoken English in business communication.	a) yes	b) no	
3	If yes, what is the frequency of using English for business communication?	a) daily	b) occasionally	
4	Your company employs business graduates majoring in one of the following: <i>Marketing, Management, Accounting and Business Finance.</i>	a) yes	b) no	
5	Your company expects business graduates to be able to use oral English skills in business communication.	a) yes	b) no	
6	When an individual speaks English, they apply ____ style(s).	a) conversational	b) presentational	c) both
7	Rank in the order of importance the following speaking tasks in English (1 being the most important, 7 being the least important):	informal conversation _____ phone call _____ group presentation _____ interview _____ giving a statement – formal (e.g. PR) _____ chat with colleagues _____ providing explanation/description (short monologue) _____		

Questions 8 to 21 are provided in the second part of the questionnaire, titled as Business presentations (Appendix I). The questions in this part focus on presentational skills, the setting and audience, allocated time for the performance, and indigenous assessment criteria. Questions 1 to 6 are multiple choice questions, 8 of which are dichotomously scored as true/not true (or yes/no), or in one case “daily vs. occasionally”. Eight questions (more specifically, questions: 1, 6, 8, 9, 10, 11, 12 and 13) follow the multiple-choice format, but one with three options (with “option c” stating that “both a and b” are possible). Questions 18 and 19 are ranking questions

with three options, prompting the respondent to rank them in order of importance (1 being the most important, 3 being the least important). Given that the survey takes the form of an interview, the researcher’s role is to explain the prompt and ranking system to the respondents. The last two questions (20 and 21) are open-ended, prompting the respondents to provide a less structured response. Considering the fact that all survey questions are in English, the researcher recorded all answers for the purpose clarity and correctness. The total time allocated for responding to the Context-based survey was between 25 and 45 minutes.

Table XXXX. Context-based questionnaire: Part 2 - Business presentations

8	When they present in English, your employees are expected to do it_____.	a) individually	b) in a group	c) both a and b
9	When they present in English, individuals talk for_____min.	a) less than 5	b) 5-10	c) more than 10
10	When they present in English, the presentation can take place _____.	a) live	b) via video-conference call	c) both a and b
11	In an average business presentation, the number of the people in the audience is in the following range:	a) 1-5	b) 6-10	c) more than 10
12	People in the audience are:	a) colleagues	b) business associates/clients	c) both a and b
13	The communication and setting during presentations are:	a) formal	b) informal	c) both a and b
14	In an average business presentation, the people in the audience ask questions related to the content of the presentation.	a) yes		b) no
15	While presenting, the presenter(s) is/are required to manipulate	a) yes		b) no

	equipment/use visuals/perform demonstrations.		
1 6	While presenting, the presenter(s) is expected to use technical words/specialized vocabulary? (e.g. related to the products/production/specificities of the company itself, etc.)	a) yes	b) no
1 7	When presenting in English, your employees, with educational background in economics are expected to demonstrate the knowledge they gained in university.	a) yes	b) no
1 8	Can you rank the following in the order of importance (1 being the most important, 3 being the least important in a presentation)?	self-confidence and persuasiveness _____ clear organization and structure _____ native-like pronunciation _____	
1 9	Can you rank the following in the order of importance (1 being the most important, 3 being the least important)?	grammatical accuracy _____ fluency and voice projection _____ content and technical vocabulary _____	
2 0	Can you provide examples of presenting in English (consider who the presentations is delivered for? in what setting? how long is it? are there any special materials that presenters provide?)		
2 1	Can you provide any criteria by which you judge the success of a presentation in English?		

General context questions provide the idea of business environment and the company requirements regarding the employment of business graduates and their English language skills. Further to this, it provides insight into the frequency of English language use in business communication. Although the research investigates spoken production, it can be deduced that some type of written production in English is involved as well (presentation scripts/scenarios, PowerPoint presentations, print outs, promotion materials, etc.). Seven out of twenty-five company representatives interviewed state that their company performs business operations locally, in Serbia, meaning that they do not import/export goods and services and do not have any branch offices or affiliated companies abroad; or that their company is not a type of authorized dealership or franchise. However, they report performing activities that involve speaking English skills (attending international fairs, sending delegates to professional development programs). Most companies use spoken English on a daily basis (18 out of 25) in order to achieve various communicative goals. In addition, all the respondents report that their companies offer employment to business graduates, agreeing that they are expected to demonstrate the knowledge of English language in addition to their knowledge of economics. Further to this, the responses to Q6 reveal that communication in English involves both spoken production and interaction (15 out of 25 respondents agree that their employees employ both conversational and presentational style when speaking English), with only 3 cases restricted to conversational English, whereas 7 out of 25 report employing presentational style only. Question 7 reveals the respondents' opinion regarding relevance of certain speaking task types requiring that speakers use English while performing on them. The ability to participate in a group presentation in English and the skill of sustaining a short monologue are considered as highly relevant skills. They are followed by the skill of making a phone call, participating in an interview in English (both predominantly interactive tasks), and giving a formal statement. Informal conversation and chat are considered the least important in a business setting (Table 7.2).

Table 7.2. Results of the survey: responses to the General part of the Context-based questionnaire (questions 1-6)

1	a: 7 b: 0 c: 18														
2	a: 25 b: 0														
3	a: 18 b: 7														
4	a: 25 b: 0														
5	a: 25 b: 0														
6	a: 3 b: 7 c: 15														
7	<table> <tr> <td>group presentation</td> <td>1</td> </tr> <tr> <td>providing explanation/description (short monologue)</td> <td>1</td> </tr> <tr> <td>phone call</td> <td>2</td> </tr> <tr> <td>interview</td> <td>3</td> </tr> <tr> <td>giving a statement – formal (e.g. PR)</td> <td>4</td> </tr> <tr> <td>informal conversation</td> <td>5</td> </tr> <tr> <td>chat with colleagues</td> <td>6</td> </tr> </table>	group presentation	1	providing explanation/description (short monologue)	1	phone call	2	interview	3	giving a statement – formal (e.g. PR)	4	informal conversation	5	chat with colleagues	6
group presentation	1														
providing explanation/description (short monologue)	1														
phone call	2														
interview	3														
giving a statement – formal (e.g. PR)	4														
informal conversation	5														
chat with colleagues	6														

The insight into business activities provided in the General context questionnaire help test developers by understanding the context in which, business graduates perform various duties and job responsibilities, and are expected to use spoken English for business communication on a daily basis. In addition to this, in most cases, the participants in TLU situations are expected to participate both in presentational and conversational speaking events (in 15 out of 25 cases). Table 7.2 above presents the summary of responses to the General part of the questionnaire.

7.3.1.2 From general context to specific tasks

The second part of the questionnaire aims at eliciting more concrete responses that reveal the specific purpose target language use situations in which individuals and groups use spoken English language to achieve their respective communicative goals. The findings demonstrate that two most frequently occurring tasks refer to an individual and/or a group presentation, since they exert a significant impact on their company’s business operations. This part of the questionnaire indicates that business presentations take place live and in predominantly formal settings (Q10 and Q13), and are attended by 6 to 10 people who can be both business associates and clients, although they are sometimes attended by colleagues in the role of audience (Q11 and Q12).

Individual presenters deliver presentations in English, speaking from 5 to 10 minutes each, and are expected to demonstrate their background knowledge as well as the knowledge of technical vocabulary. While presenting, the individuals are expected to operate electronic equipment (laptop, projector, presentation pointer); demonstrate how products are used or how services are performed; use and interpret visuals (graphs, charts and tables) (Q9, Q15, Q16, and Q17). People in the audience normally expect some sort of audience engagement and ask questions related to the content of the presentation (Q14). The responses to questions 17 and 18 demonstrate the indigenous assessment criteria (or the assessment criteria applied by the participants in a communicative act), indicating that subject specialist informants value clear organization and structure of the presentation more than what can be described as a native-like pronunciation. Demonstrating self-confidence and structuring the presentation well is perceived as more important than sounding like a native speaker. In non-native settings, the performance is often judged against the performance of native speakers; however, Luoma argues that very few learners can achieve the native-like standard in all aspects of their performance, adding that native speakers' performance is "so varied that it can hardly be taken for a standard" (2004:10). Another important finding refers to the evaluation of performance where the content and technical vocabulary take precedence over grammatical accuracy. Fluency and voice projection seem to be more important than grammatical accuracy as well. This reflects Douglas's findings confirming that specialist subject informants add more value to the achievement of communicative goals than to grammatical accuracy of their performance (2000). In learning settings, however, instructors must devise means for reconciling the two, since grammatical knowledge is something that they need to teach and test in order to achieve learning objectives set by the course syllabus. The table below shows the summary of responses to the Business presentation part questionnaire (Table 7.3).

Table 7.3 Results of the survey: responses to the Business presentation questionnaire (questions 8-19)

8	a: 7 b: 8 c: 10
9	a: 3 b: 20 c: 2
10	a: 16 b: 6 c: 3
11	a: 2 b: 20 c: 3
12	a: 4 b: 1 c: 18
13	a: 2 b: 3 c: 0
14	a: 2 b: 0
15	a: 2 b: 3
16	a: 25 b: 0
17	a: 20 b: 5
18	1 clear organization and structure (11) 2 self-confidence and persuasiveness (10) 3 native-like pronunciation (4)
19	1 content and technical vocabulary (10) 2 fluency and voice projection (9) 3 grammatical accuracy (6)

The last two questions in the questionnaire follow the open-ended format, allowing the respondents more freedom in answering. The responses to question 20 reveal TLU situations and actual real life tasks taking place in them, whereas the responses to question 21 give insight into rating procedures and criteria for correctness (as a form of indigenous assessment criteria) by revealing some of the indicators against which the success of a communicative act is evaluated (see summarized responses in Table 7.4 below). The distinction is made between the indicators that can be applied *immediately* and those that can be applied *subsequently*, or after the communicative event. The immediate indicators of success include some of the following: immediate expressions of satisfaction, the purchase of goods and services, asking follow-up questions, expressing interest in the topic, whereas the subsequent indicators include the positive reactions following the communicative event. These notions of successful performance, being

indicators of indigenous assessment criteria, are crucial for developing rating scales that are used in assessing performance on test tasks.

Table 7.4 Results of the survey: responses to the Business presentation part of the Context-based questionnaire - summary (questions 20-21)

20	<ul style="list-style-type: none"> - individual/ group presentation of a product / service (for business associates, prospective clients/ existing client) - individual/group presentation of the company (its mission, vision, range of products/service/ future plans for expansion/new markets (at fairs, exhibitions, joint presentations, Chamber of Commerce events, cluster events, etc.) - project/service demonstration (usually performed by an individual) for prospective client(s) focusing on pros and cons and the company’s relative standing in comparison to the competitors (price, quality, maintenance, warranty duration, extra services, etc.) - project launching (group presentation) at a fair/in-house exhibition/TV show/ Internet-streaming / Instagram TV/live - video-conference call and presentation of a product/service/ research and development results - Questions and Answers (Q&A) sessions – addressing questions/issues/resolving problems/defending a product/solution/ service in groups, pairs, and individually 	
21	<p><u>immediate:</u></p> <ul style="list-style-type: none"> - immediate expression of satisfaction (customers, business associates, colleagues, managers, etc.) - product/service orders/purchases - follow-up questions asked - questions expressing interest raised - positive feedback received during the presentation (live/Instagram posts/ instant messages/phone calls to the company headquarters/ hot-line, live feed, etc.) 	<p><u>subsequent:</u></p> <ul style="list-style-type: none"> - (e-mail/ phone call/ Viber/ Whatsapp/ Instagram post, instant text) messages expressing satisfaction (clients, business associates, colleagues, managers, etc.) - contracts/agreements renewed - products/services commissioned - letters of interest received from prospective/existing clients - customer satisfaction surveys showing positive results

The responses to the questionnaire provided by company representatives provide valuable insight into TLU situations and tasks. Their commentaries related to the task types, task achievement and the criteria by which the success in performance is evaluated will form a basis for creating speaking test tasks which will be used in educational domain. To ensure that test tasks and TLU tasks share the same characteristics, the TLU tasks will be analyzed by using Task Characteristics Framework (See Chapter 7.4.1 below), and then test tasks will be developed based on the characteristics of the TLU tasks. It is the author’s assumption that newly created tasks will be (situationally) authentic and relevant to the real life domain, ensuring students’

engagement and interaction with the task characteristics in the manner language users interact with the task in the real life domain (interactional authenticity).

7.4 Relating TLU speaking tasks to speaking test tasks

In this chapter, the author will use the Task characteristics framework to analyze TLU speaking tasks, based on the responses received from the subject specialist informants. Following the analysis, the author will analyze speaking test tasks, providing test task specifications for the use within the educational domain in Phase 2 of the research.

7.4.1 TLU task characteristics

Given the importance that authenticity has in communicative language testing, and, consequently, in specific purpose language testing, language testers have to devise the ways to ensure that test tasks and test takers' interaction with the tasks resemble the TLU situation. Once the TLU situations have been analyzed, test developers proceed by analyzing language tasks occurring in them. As outlined in Chapter 5.2.1 above, TLU language tasks are worth considering as their characteristics are used for: (1) modeling test tasks, (2) enabling the engagement of test takers' language ability, and (3) determining test task authenticity (both in situational and interactional sense) and investigating the validity of inferences based on test scores (Bachman, 1990; Bachman and Palmer, 1996; Douglas, 2000; Chapelle and Douglas, 2006). To this end, the author will use the Task characteristics framework to analyze target language use situations and develop speaking test tasks that share the same characteristics. The Task characteristics framework used here is based on the framework that Douglas suggests that test developers should use in the context of testing languages for specific purposes (see Chapter 5.2.1). The following sets of task characteristics will be used to analyze the data collected in collaboration with subject specialist informants:

- (1) the rubric,
- (2) the input,
- (3) the expected response,
- (4) the interaction between the input and response, and
- (5) the assessment.

Having interviewed the subject specialist informants, the author came to the conclusion that speaking tasks fall into two major categories: presentational and conversational tasks. The former are considered by the subject specialist informants to be “more important”, requiring better developed speaking and presentational skills (see Chapter 7.3.1.1 above). For this reason, the following discussion of TLU tasks focuses on extended oral production tasks delivered in English. More precisely, it delineates two particular task types that are prominent in TLU situations:

- 1) a group speaking task (a presentation in English)
- 2) an individual speaking task (a short individual presentation in English)

7.4.1.1 Group speaking task (presentation) – TLU task characteristics

7.4.1.1.1 The rubric

The characteristics of task rubric specify how language users are supposed to react and use their language skills in a particular situation. The following are characteristics grouped within this set: objective, procedures for responding, structure, time allotment and evaluation (see discussion in 5.2.1.1 above). It should be noted that the majority of these characteristics are quite implicit, embedded in the communicative situation. In a testing context, however, they need to be made explicit so that test takers know how they should attend to the task. The characteristics of the rubric for the group speaking task in the TLU indicate that language users participate in a joint venture of delivering a presentation to the audience interested in their company’s product, service, or the company’s activities and plans. The event usually lasts for at least 30 minutes, depending on the purpose of the message that is to be conveyed.

Table 7.5. Characteristics of target language situation tasks – group speaking task (the rubric)

Characteristics of the rubric	
Specification of the objective	Implicit in TLU situations: as a member of a group prepare and deliver a talk about a product/service/company to the audience comprising existing or prospective clients
Procedures for responding	Implicit: prepare the talk well in advance in English, rehearse it, and deliver it (orally) to an audience by using visuals
Structure of the communicative event	
Number of tasks	One complex task (involves the preparation and delivery)
Time allotment	30+ min. presentation
Evaluation	
Criteria for correctness	Implicit: the expression of customer satisfaction, follow-up questions, placement of orders for goods and services
Rating procedures	<u>Implicit and informal</u> : embedded in the communicative event. <u>Explicit and formal</u> : Supervising managers observe, evaluate and provide feedback to the participants in the event in question.

The evaluation criteria are implicit, embedded in the context, with audience responding to the presentation and taking appropriate follow-up actions. When it comes to rating procedures, they can be implicit or explicit, in the form of superiors' observations, feedback and follow-up actions (Table 7.5).

7.4.1.1.2 The input

The input's role is to ensure that language users have enough contextual cues to respond to situational tasks appropriately. By analyzing the characteristics of the prompt, language users analyze contextual information (pertaining to the setting, participants, purpose, and the form and content of the prompt) helping them assess the context, employ the appropriate strategies for responding, and execute the response. The input data characteristics, on the other hand, refer to the materials given to language users to process them and respond accordingly. In the group speaking task, language users may find themselves in various settings, with various participants, and for different reasons. However, regardless of many differences, the Table 7.6 outlines the characteristics that the majority of identified TLU situations share.

Table 7.6. Characteristics of target language situation tasks – group speaking task (the input)

Characteristics of the input	
Prompt	
Features of the LSP context	
Setting	The settings vary significantly from one context to another. Some of the shared characteristics are as follows: conference room/ presentation hall, chairs or designated space for the audience (could be outdoors as well); laptop/desktop computer, projector, presentation pointer/clicker; loudspeakers, microphones; may involve media coverage (cameraman, reporter) and a photographer; promo material/handouts/ samples;
Participants	Three to five presenters and the audience of more than 5 people. The people in the audience can vary regarding their respective roles and expectations (clients, spectators, general audience, business partners, associates, competitors, etc.). Ethnically heterogeneous people, male and female, in all age groups. Usually not very familiar to the presenter(s), except in the case of long-standing business partners.
Purpose	Purposes vary from context to context. Some of the most recurring purposes include but are not limited to the following: <ul style="list-style-type: none"> - to promote a product / service together with other group members (in front of: business associates, prospective clients/ an existing client) - to act as a member of a group and give an overview of the company’s mission, vision, range of products/service/ future plans for expansion/new markets (at fairs, exhibitions, joint presentations, Chamber of Commerce events, cluster events, etc.) - to launch a product/service(group presentation) at a fair/in-house exhibition/TV show/ Internet-streaming / Instagram TV/live - to participate in a video-conference call and presentation of a product/service/ research and development results - to attend Questions and Answers (Q&A) sessions – addressing questions/issues/resolving problems/defending a product/solution/ service together with colleagues
Form and content	
Tone	Businesslike, varying degrees of formality and friendliness; persuasive
Language	World English with varying degrees of a foreign accent
Norms of interaction	Presenters/audience interaction; colleague/colleague interaction; business representatives/client interaction

Genre	Presentation
Problem to address	Implicit in the TLU: to show the benefits of a product/service. To provide detailed account of the company's plans and activities (including their mission and vision)
Input data	
Format	
Visual	Written/audio/video material in the PowerPoint/Prezzi presentation; printed promotional material; manuals and instructions
Audio	Original or copyrighted audio recordings, oral questions from the audience
Vehicle of delivery	Live; oral, written
Length	In the range from a couple of hours to a couple of days to process the input data
Level of authenticity	
Situational	By definition
Interactional	Deeply engaged

The characteristics pertaining to the input data in TLU situations include the format, vehicle of delivery, length and the level of authenticity (Table 7.6 above). Bearing in mind that when it comes to an oral presentation, the input data may come in different formats depending on the source and the situation, the input data are authentic by definition. They not only include authentic texts (spoken or written) that language users process, but language users' interaction with such texts is authentic as well.

7.4.1.1.3 The expected response

The characteristics pertaining to the expected response in a TLU are related to participants' expectations related to other participants' reactions and responses. In a testing situation, this set of characteristics refers to what assessors expect that test takers should do after they have processed the input – react physically, select an option in a MCQ test format, etc. Having analyzed the TLU situation, the author came to the findings summarized in Table 7.7 below.

Table 7.7. Characteristics of target language situation tasks – group speaking task (the expected response)

Characteristics of the expected response	
Format	
Written	May include visuals and printed material
Oral	Extended oral production involving the use of visuals (including project demonstration)
Physical	Product operation, demonstration in front of the audience (explaining processes), manipulating equipment
Type of response	
Selected	
Limited production	
Extended production	Extended response (over 5 min)
Response content	
Nature of language	Vocabulary appropriate to the topic and audience
Background knowledge	Topic-related knowledge, economics, corporate culture norms, familiarity with the culture the audience comes from
Level of authenticity	
Situational	Contextualized, it is being built on the spot as the presentation evolves
Interactional	Deeply engaged

As can be seen, group oral presentations may include various formats of responses, including the presentation of written/printed material, viewing and discussing various sources of visuals, and demonstration of a product or a process. Given that the identified genre is oral presentation, the response takes the form of an extended spoken production, characterized by topic-appropriate (often highly specialized) vocabulary. The authenticity pertaining to language users' response is embedded in context, with language users deeply engaged in the task.

7.4.1.1.4 The interaction between the input and response

This set of characteristics describes the nature of the relationship between the input and the expected response, showing how much the response depends on the input. Group presentations normally take place in front of the audience that interacts with presenters. Depending on the audience's reaction to the input provided by presenters, the presenters may adapt and modify their narrative, accommodating all the requests for clarification or additional information coming from the audience or other members of the presenting group.

Table 7.8. Characteristics of target language situation tasks – group speaking task (the interaction between the input and response)

Characteristics of the interaction between the input and response	
Reactivity	
reciprocal: non-reciprocal	On the continuum from somewhat reciprocal to fairly reciprocal, depending on the feedback the presenters get from the audience
Scope	
broad-narrow	Very broad
Directness	
dependent upon input: dependent upon background knowledge	On the continuum from somewhat direct to fairly indirect (as the speakers have to process some information from the input, but they also rely on their background knowledge in attending to the task).

As can be concluded from the summary of findings presented in Table 7.8, the reactivity of the interaction is set on a continuum from somewhat reciprocal to fairly reciprocal, while the input that has to be processed for this task type involves a very broad scope of interaction. In other words, language users need to process a lot of input material to prepare their responses (individual contributions to the task). The characteristic of directness investigates the dependence on the background or topical knowledge in response to the task. In the case of a business presentation, this relationship can take any place on the continuum from somewhat direct to fairly indirect.

7.4.1.1.5 The assessment

The assessment characteristics are those related to defining the language ability necessary to execute the task, the criteria for correctness, and rating procedure. The language ability in TLU situations is quite complex (see the summary, Table 7.9). The criteria for correctness reveal the indigenous criteria, or what participants in a situation consider as correct (or sufficient). Rating procedures are quite interesting, indicating that the manner in which a performance is “assessed” normally comes with a result. The results can be implicit and immediate (positive reactions, follow-on and follow-up questions, expressing interest, etc.) or they can be explicit and immediate/subsequent (purchasing orders, letters of interest, feedback, etc.).

Table 7.9. Characteristics of target language situation tasks – group speaking task (the assessment)

Characteristics of the assessment	
Construct definition	Specific purpose language ability is quite complex in the TLU situations in the observed TLU domain. Some of the shared characteristics are as follows: general business terminology, the knowledge of marketing terminology and customer relations norms; pan-technical terminology; the use of declaratives, tag questions and rhetorical questions, indirect and Wh-questions; the cohesive use of discourse markers; organization knowledge of process structure, transitions and turn-taking strategies; use of heuristic, ideational, and manipulative functions; common idioms and cultural references; strategic use of presentational style, the ability to operate devices and manipulate various pieces of digital equipment (computers, projectors, etc.) involving audience by asking them questions or involving them by short and hands-on activities; using comprehension checks. Background knowledge: ability to elaborate on the topic by using the terminology everyone in the audience is likely to understand; awareness of presentational conventions.
Criteria for correctness	Indigenous criteria: presentation skills, pronunciation and comprehensibility, voice projection, cultural awareness, content/background knowledge, presenters' personality and experience (friendly, professional, responsive, knowledgeable)
Rating procedures	Implicit and immediate: the members of the audience assess the presenters informally by means of their questions, comments, purchasing orders, follow-up activities or questions; Explicit and subsequent: purchasing orders, emails and messages expressing (dis-)satisfaction, follow-up questions and activities Explicit and (normally) subsequent: supervisors and managers assess the success of the event (presentation) by feedback/promotion/sanctioning

7.4.1.2 Individual speaking task (short talk/ presentation) – TLU task characteristics

7.4.1.2.1 The rubric

The characteristics of rubric in an individual presentation TLU task include the following: objective, procedures for responding, structure, time allotment and evaluation. In the TLU situation, these characteristics are implicit, set by the context, with participants relying on their strategic competence and the knowledge of the context when making their strategies for

responding. In a test, the characteristics of the rubric are made explicit. Table 7.10 outlines the summary of the task rubric characteristics that the analyzed individual TLU tasks share.

Table 7.10. Characteristics of target language situation tasks – individual speaking task (the rubric)

Characteristics of the rubric	
Specification of the objective	Implicit in TLU situations: to deliver a short talk expressing opinions (by explaining, describing, justifying, demonstrating, instructing) about a certain problem/situation/issue to an interlocutor (or small audience)
Procedures for responding	Implicit: interact orally in English, explaining own point of view in a short monologue
Structure of the communicative event	
Number of tasks	Varies by the number of questions asked/ one task involving the preparation of an answer
Time allotment	3-5 minutes
Evaluation	
Criteria for correctness	Implicit: the interlocutor's satisfaction with provided argumentation/ problem or issue resolved/ the response addresses the issue in its entirety
Rating procedures	Implicit: embedded in the communicative event. The interlocutor responds to the talk, stating their (dis-)satisfaction with the argumentation/explanation.

It can be noted that participants engage in communication to meet various communicative purposes/functions (explain, compare/contrast, describe, justify, persuade, demonstrate, etc.). The number of tasks may vary in a TLU situation, depending on the reactivity of the interaction between the input and expected response. The evaluation characteristics are implicit and embedded in context, with participants demonstrating their own (dis-)satisfaction with the response/ the manner in which a problem is being handled.

7.4.1.2.2 The input

When it comes to the input characteristics pertaining to individual speaking tasks in TLU situation, it should be noted that the prompt characteristics are often implicit and highly

contextualized. The participants in the speaking task are aware of the setting and other participants and are focused on the purpose of the task. The following recurring purposes for an individual presentation/talk have been identified in the TLU situation: talking about a product/service for promotional purposes; talk about the company’s mission, vision, plans; describe how something works; address issues of various kinds, providing explanations, justifications, and assistance. The prompt characteristics reveal that individual speaking tasks do not last long; they require limited processing of input data, and the tone of the speaker is important (professional, helpful, restrained, and friendly). The problem that needs to be addressed is implicit in the target language use situation, and it involves the participants whose norms of interaction are on the continuum from casual to formal (see Table 7.11 below).

Table 7.11. Characteristics of target language situation tasks – individual speaking task (the input)

Characteristics of the input	
Prompt	
Features of the LSP context	
Setting	The settings vary from one context to another. Some of the shared characteristics are as follows: office/ business premises; table or a booth, could be office cubicle as well; printed material/ various objects/ computer screen, Internet connection.
Participants	One person in the role of the speaker, usually no more than 2-3 other people who listen to the talk. Usually unfamiliar to the speaker, people come to ask explanation/solution to the problem, seek advice or are curious about the description provided by the speaker.
Purpose	<p>Purposes vary from context to context. Some of the most recurring purposes include but are not limited to the following:</p> <ul style="list-style-type: none"> - to (individually) promote a product / service (in front of: business associates, prospective clients/ an existing client) - to (individually) give an overview of the company’s mission, vision, range of products/service/ future plans for expansion/new markets (at fairs, exhibitions, joint presentations, Chamber of Commerce events, cluster events, etc.) - to demonstrate how a product works/ how a service is provided for prospective client(s) focusing on pros and cons and the company’s relative standing in comparison to the competitors (price, quality, maintenance, warranty duration, extra services, etc.) - To address questions/ complaints/ and various

	issues that may arise. To defend a product/service/solution in front of consumers and other interested parties
Form and content	Time-limited language production provided by an individual in a question/answer format
Tone	Friendly and professional; restrained and fair; analytical
Language	World English with varying degrees of a foreign accent
Norms of interaction	Business representative/colleague/client/interested parties interaction. On the continuum from casual to formal
Genre	Question/answer session; answer in the form of a short monologue-like presentation
Problem to address	Implicit in the TLU: <ul style="list-style-type: none"> - to address various issues/problems/complaints/ - to explain/describe/justify various aspects of a product/service/situation/activity
Input data	
Format	
Visual	Written instructions/ directions; product specifications/ portfolio/ rules and procedures/ terms and conditions/ manuals
Audio	Questions from the interlocutors (clients, colleagues, business partners)
Vehicle of delivery	Live; oral; written
Length	1-2 (or more) hours to study the input data
Level of authenticity	
Situational	by definition
Interactional	Engaged (on a continuum from somewhat engaged to deeply engaged)

The input data are characterized by their format and the level of authenticity. As is the case with the group speaking task, the materials that language users have to process can take the form of a written text (instructions, manuals, portfolios, product specifications, various written documents); audio-video recording (tutorials, instructions, recorded message); the aural input from other participants, etc. (see Table 7.11 above). The situation itself is authentic by definition, because it occurs within the “natural” context, with participants engaged into setting targets and achieving their communicative goals as per the situational cues.

7.4.1.2.3 *The expected response*

As is the case with testing contexts, in the real life contexts, participants respond to contextual cues and respond to stimuli engaging their pragmatic and language knowledge, and

depending on the demands of a situation, they rely on their background knowledge to achieve particular communicative goals. The expected response is characterized by its format, type, content, and the level of authenticity. In an individual speaking tasks, language users provide their responses orally (sometimes accompanied by a live demonstration or writing) by employing the ability to produce extended speech. The length of the speech may vary depending on the needs of a particular situation, and the language function that is being performed (explanation, description, etc.)

Table 7.12. Characteristics of target language situation tasks – individual speaking task (the expected response)

Characteristics of the expected response	
Format	
Written	May include visuals and printed material
Oral	Oral explanation of the problem, sometimes accompanied by demonstrating how something works
Physical	Manipulating a piece of equipment
Type of response	
Selected	
Limited production	
Extended production	Extended response (5-10 min)
Response content	
Nature of language	Vocabulary appropriate to the topic and audience
Background knowledge	Topic-related knowledge, economics, corporate culture norms
Level of authenticity	
Situational	Building on the problem stated by the interlocutor, the presenter responds to the best of their knowledge. Situational authenticity is embedded in the context
Interactional	Deeply engaged

In specific purpose target language situation the content of the language employed is characterized by specific purpose vocabulary. Language users employ their language knowledge, strategic competence and background knowledge to respond to the demands of a situation.

Situational authenticity is embedded in the context, urging participants in the communicative act to engage in the task (Table 7.12).

7.4.1.2.4 The interaction between the input and response

Individual speaking tasks taking place in a target language use situation normally address a purpose for speaking. The participants engage in the task in order to solve a problem, describe a process, and resolve an issue. This implies that the interaction between the input and the expected response is highly reciprocal, with each participant adapting to the previous utterance of the interlocutor. As for the scope and directness of the interaction, it is very broad and direct, as the language users process a lot of information from different sources and rely on the input to attend to the task (resolve an issue with their clients). However, the task completion may require that background knowledge be employed, which places the interaction on the continuum from indirect to fairly direct (Table 7.13).

Table 7.13. Characteristics of target language situation tasks – individual speaking task (the interaction between the input and response)

Characteristics of the interaction between the input and response	
Reactivity	
reciprocal: non-reciprocal	Highly reciprocal (all parties need to adapt as necessary so as to ensure mutual comprehension)
Scope	
broad-narrow	Very broad
Directness	
dependent upon input: dependent upon background knowledge	It can be anywhere on the continuum from indirect to fairly direct, depending on the following: <ul style="list-style-type: none"> a) the speaker can attend to the task by relying on their background/topical/technical knowledge, b) the speaker has to rely on the input (e.g. equipment manual) to respond. or c) both

7.4.1.2.5 The assessment

The assessment characteristics reveal the construct underlying the ability to perform a language task in a TLU situation, criteria for correctness in a particular situation, and rating criteria employed to perform the assessment. Specific purpose language ability required to perform an individual speaking task is quite complex. The individual engaged in the task has to possess relevant linguistic, pragmatic and background knowledge to cater for the specificity of the situation. The summary of the component parts of the construct for this task is outlined in Table 7.14.

Table 7.14. Characteristics of target language situation tasks – individual speaking task (the assessment)

Characteristics of the assessment	
Construct definition	Specific purpose language ability is quite complex in the TLU situations in the observed TLU domain. Some of the shared characteristics are as follows: general business terminology, customer relations (including customer support) terms, the knowledge of marketing and pan-technical terminology; the use of declaratives, tag questions and rhetorical questions, indirect and Wh-questions; the cohesive use of discourse markers; organization knowledge of process structure, transitions and turn-taking strategies; use of heuristic, ideational, and manipulative functions; common idioms and cultural references; problem-solving skills; strategic use of expository and conversational styles, using comprehension checks. Background knowledge: ability to elaborate on the topic by using the terminology the client/interlocutor can easily understand; awareness of service provider/client business conventions
Criteria for correctness	Indigenous criteria: presentation skills, pronunciation and comprehensibility, voice projection, cultural awareness, content/background knowledge, presenters' personality and experience (friendly, professional, responsive, learned)
Rating procedures	Implicit: the interlocutor(s) may make private judgments of the communicative act, or they can take a more active approach and ask questions, make/withdraw the purchase/order; expressions of (dis-)satisfaction Explicit: supervisors and managers assess the success of the communicative event by feedback/promotion/sanctioning

Criteria for correctness reveal what participants in the TLU situation consider as correct (or sufficient) response: presentation skills, pronunciation and comprehensibility, voice

projection, cultural awareness, content/background knowledge, presenters' personality (friendly, professional, responsive, learned) and experience. At the same time, criteria for correctness provide a basis for rating the speakers' performance. The rating can be both implicit and explicit, depending on the situation and the role of other participants (the summary of rating procedures is provided in Table 7.14 above).

7.4.2 Test task characteristics

The following step in the transition from real life domain to the domain of education is to analyze prospective test tasks and determine the extent to which they correspond to TLU tasks. The same Task characteristics framework is applied to compare TLU tasks to test tasks, and provide the basis for test task specifications (Chapter 7.4.3). The analysis of the TLU situations helped the author identify two recurring speaking tasks:

- a group speaking (test) task, and
- a short individual speaking (test) task.

7.4.2.1 *Group speaking task*

In this chapter, the author utilizes the Task characteristics framework to analyze the characteristics of the group speaking task (group presentation). The findings will be used in Chapter 7.4.3 to develop test task specifications for actual use in the educational domain (Phase 2 of the research). The following tables summarize the test task analysis, showing how the Task characteristics framework can be used as a Test task characteristics framework with the purpose of comparing TLU tasks to potential test tasks (see the Tables 7.15- 7.19 below).

7.4.2.1.1 *The rubric*

The following table contains the characteristics of the rubric, comparing the rubric of a TLU speaking task to that of a test task (see Table 7.15 below). The task in question is performed in a group.

Table 7.15. Comparison of task characteristics of the TLU and test – group speaking task (the rubric)

Characteristics	TLU situation	Test task
Rubric		
Specification of the objective	Implicit in TLU situations: as a member of a group prepare and deliver a talk about a product/service/company to the audience comprising existing or prospective clients	Explicit: to assess English oral ability in the context of a group business presentation
Procedures for responding	Implicit: prepare the talk well in advance in English, rehearse it, and deliver it (orally) to an audience by using visuals	Explicit: work in a group to collect data and prepare a 10-15 min. long presentation (including visuals) on a chosen topic and deliver it in English
Structure of the communicative event		
Number of tasks	One complex task (involves the preparation and delivery)	One task requiring a thorough preparation
Time allotment	30+ min presentation	Phase 1: 4 weeks to collect data and prepare for the presentation Phase 2: 10-15 minutes for the delivery
Evaluation		
Criteria for correctness	Implicit: the expression of customer satisfaction, follow-up questions, placement of orders for goods and services	Explicit: overall comprehensibility; communication skills (organization and presentation); interaction with the audience; overall impression
Rating procedures	Implicit and informal: embedded in the communicative event. Explicit and formal: supervising managers observe, evaluate and provide feedback to the participants in the event in question.	Explicit and formal: Two raters use analytic/holistic rating scale to score performance independently (ratings averaged); analytic: 4 categories scored on a scale of 1-5; holistic: scores on a scale of 1-10;

7.4.2.1.2 The input

The following table summarizes the comparison between the input used in the TLU situation for a group speaking task to the more explicit rubric that must be developed for a speaking assessment (see Table 7.16 below).

Table 7.16. Comparison of task characteristics of the TLU and test – group speaking task (the input)

Characteristics	TLU situation	Test task
Input		
Prompt		
Features of the LSP context		
Setting	The settings vary significantly from one context to another. Some of the shared characteristics are as follows: conference room/ presentation hall, chairs or designated space for the audience (could be outdoors as well); laptop/desktop computer, projector, presentation pointer/clicker; loudspeakers, microphones, may involve media coverage (cameraman, reporter) and a photographer; promo material/handouts/ samples;	A theater with a stage and podium; blackboard, flip-chart and flip-chart holder, an overhead projector connected to a desktop computer (with the Internet connection) and a large screen above the blackboard, a presentation laser pointer/PowerPoint clicker; lapel and hand microphones (wireless); rows of seats and computer desks; lights, small side windows, and the AC-controlled room temperature; a large table between the blackboard and the audience (may hold the exhibits) the theater with the seating capacity of 300 in the audience; the presenters told to reveal the purpose of the presentation (e.g. launching a new product) to help the audience imagine the setting
Participants	Three to five presenters and the audience of more than 5 people. The people in the audience can vary regarding their respective roles and expectations (clients, spectators, general audience, business partners, associates, competitors, etc.). Ethnically heterogeneous people, male and female, in all age groups. Usually not very familiar to the presenter(s), except in the case of long-standing business partners.	In each session there are 25 students in the audience including the presenting group; male and female students (aged 20-30); two instructors seating unobtrusively in the audience
Purpose	Purposes vary from context to context. Some of the most recurring purposes include but are not limited to the following: <ul style="list-style-type: none"> - to promote a product / service together with other group members (in front of: business associates, 	Assessment of English ability to deliver a group oral presentation on a business topic

	<p>prospective clients/ an existing client)</p> <ul style="list-style-type: none"> - to act as a member of a group and give an overview of the company’s mission, vision, range of products/service/ future plans for expansion/new markets (at fairs, exhibitions, joint presentations, Chamber of Commerce events, cluster events, etc.) - to launch a product/service (group presentation) at a fair/in-house exhibition/TV show/ Internet-streaming / Instagram TV/live - to participate a video-conference call and presentation of a product/service/ research and development results - to attend Questions and Answers (Q&A) sessions – addressing questions/issues/resolving problems/defending a product/solution/ service together with colleagues 	
Form and content	Presentation including questions from the audience	A group presentation in front of an audience
Tone	Businesslike, varying degrees of formality and friendliness; persuasive	Persuasive and businesslike, friendly towards the audience
Language	World English with varying degrees of a foreign accent	English language (regardless of variety and foreign accent)
Norms of interaction	Presenters/audience interaction; colleague/colleague interaction; business representatives/client interaction	Presenters/audience interaction
Genre	Presentation	Business presentation
Problem to be addressed	Implicit in the TLU: to show the benefits of a product/service. To provide detailed account of the company’s plans and activities (including their mission and vision)	Explicit: <ul style="list-style-type: none"> - to provide a comprehensive and interesting account related to a company of students’ own choice (its mission, vision, operations, plans, etc.) - to use a persuasive language and

		present on a project/service of students' own choice
Input data		
Format		
Visual	Written/audio/video material in the PowerPoint/Prezzi presentation; printed promotional material; manuals and instructions	Written material coming from various sources (company website, printed promotional materials, product portfolio, student research)
Audio	Original or copyrighted audio recordings, oral questions from the audience	Audio/video recordings; questions from the audience
Vehicle of delivery	Live; oral	Live; oral
Length	In the range from a couple of hours to a couple of days to process the input data	It may vary; it should provide students with enough material (optional-in the form of a script) to sustain extended oral production (2-3 minutes per person)
Level of authenticity		
Situational	by definition	Shares many features of a TLU situation – high situational authenticity
Interactional	Deeply engaged	Students engaged in the presentation in a similar manner (but for a different purpose) as the participants in a TLU situation – high interactional authenticity

7.4.2.1.3 *The expected response*

Some of the characteristics of the expected response are deeply contextualized and embedded in a TLU situation. However, in a testing situation test takers need to know how to respond to the task prompt. The following table summarizes the analysis between the characteristics of the expected response in a TLU and in a test task (see Table 7.17 below).

Table 7.17. Comparison of task characteristics of the TLU and test – group speaking task (the expected response)

Characteristics	TLU situation	Test task
Expected response		
Format		
Written	May include visuals and printed material	Includes visuals (optional – handouts)
Oral	Extended oral production involving the viewing of visuals (including project demonstration)	Oral presentation
Physical	Product operation, demonstration in front of the audience, manipulating equipment	Manipulating a piece of equipment/ refer to visuals
Type of response		
Selected		
Limited production		
Extended production	Extended response (over 5 min)	Extended oral production (longer than 1 minute)
Response content		
Nature of language	Vocabulary appropriate to the topic and audience	Vocabulary appropriate to the topic and audience
Background knowledge	Topic-related knowledge, economics, corporate culture norms, familiarity with the culture the audience comes from	Topic-related knowledge, vocabulary related to the field of <i>Marketing</i> , Business English vocabulary covered by the course syllabus
Level of authenticity		
Situational	Contextualized, it is being built on the spot as the presentation evolves	The setting shares many features of the TLU setting allowing for situational authenticity to be on the continuum from moderate (in cases when TLU situational characteristics differ to a large extent to those of the testing context) to high (when the TLU and testing characteristics match to a great extent)
Interactional	Deeply engaged	Moderately to deeply engaged (unlike the TLU situation presenters, students have moderate experience in presenting in front of an audience)

7.4.2.1.4 The interaction between the input and expected response

The comparison of the two sets of characteristics of the interaction between the input and expected response point out the differences that exist in the quantity of input material that language users and test takers process respectively. At the same time, the Table 7.18 below indicates the role that background knowledge plays in special purpose language assessments on the concrete example of the group presentation test task.

Table 7.18. Comparison of task characteristics of the TLU and test – group speaking task (the interaction between input and response)

Characteristics	TLU situation	Test task
Interaction between the input and response		
Reactivity		
reciprocal: non-reciprocal	On the continuum from somewhat reciprocal to fairly reciprocal, depending on the feedback the presenters get from the audience	Moderately reciprocal: presenters may adapt message as necessary, but the audience and instructors might not
Scope		
broad-narrow	Very broad	Very broad
Directness		
dependent upon input: dependent upon background knowledge	On the continuum from somewhat direct to fairly direct (as the speakers have to process some information from the input, but they also rely on their background knowledge in attending to the task).	Fairly direct to somewhat direct: students depend on the input when preparing the presentation. More diligent students will capitalize on their background knowledge of marketing, but it can hardly be expected that all students employ a lot of marketing-related vocabulary since the language of instruction of the course in marketing is Serbian, and not English; also, since the presenters are still students, they will lack the practical knowledge that practitioners in the field gained through experience

7.4.2.1.5 The assessment

The comparison between the characteristics of assessment in the TLU context and in the test context indicates the following (Table 7.19):

1. A group speaking task in the TLU implies a very complex construct of language ability. Due to the constraints of practicality and validity, not all components of the TLU construct can be translated into the test task construct.
2. The criteria for correctness and rating procedures are predominantly implicit in the TLU situation, indicating that the TLU tasks employ the principles of the indigenous assessment in rating the speaking performance. The assessment criteria and rating procedures in assessing the speaking test task are explicit, helping test developers/assessors rate the performance by maximizing the qualities of reliability and validity throughout the process.

Table 7.19. Comparison of task characteristics of the TLU and test – group speaking task (the assessment)

Characteristics	TLU situation	Test task
Assessment		
Construct definition	Specific purpose language ability is quite complex in the TLU situations in the observed TLU domain. Some of the shared characteristics are as follows: general business terminology, the knowledge of marketing terminology and customer relations norms; pan-technical terminology; the use of declaratives, tag questions and rhetorical questions, indirect and Wh-questions; the cohesive use of discourse markers; organization knowledge of process structure, transitions and turn-taking strategies; use of heuristic, ideational, and manipulative functions; common idioms and cultural references; strategic use of presentational style, the ability to operate devices and manipulate various pieces of digital equipment (computers, projectors, etc.) involving audience by asking them questions or involving them by hands-on activities; using comprehension checks. Background knowledge: ability to elaborate on the topic by using the terminology everyone in the	Overall English language comprehensibility genre-appropriate and topic-appropriate vocabulary; <i>Marketing</i> -related vocabulary employed, spoken grammar, fluency and pronunciation; communication skills and confidence; appropriate non-verbal communication; use of transitions, persuasive language; use of discourse markers; use of visuals; interaction with the audience

	audience is likely to understand; awareness of presentational conventions.	
Criteria for correctness	Indigenous criteria: presentation skills, pronunciation and comprehensibility, voice projection, cultural awareness, content/background knowledge, presenters' personality and experience (friendly, professional, responsive, knowledgeable)	<p>20 points</p> <p>Point interpretation: 1 – unsatisfactory 2 – poor 3 – below expectations 4 – meets expectations 5 – above expectations</p> <p>Criteria:</p> <p>Group dynamics and Presentation structure (1-5) (time allotted; group organization and internal dynamics; presentation structure)</p> <p>Visuals and Audience engagement (1-5) (PowerPoint presentation and other visuals, relevance, imagery, audience engagement)</p> <p>Non-verbal communication (1-5) (expressiveness, confidence, non-verbal persuasiveness – posture/gestures)</p> <p>Verbal communication (1-5) (voice projection; spoken grammar; topic appropriate vocabulary; <i>Marketing</i>-related vocabulary; persuasiveness)</p>
Rating procedures	<p>Implicit and immediate: the members of the audience assess the presenters informally by means of their questions, comments, purchasing orders, follow-up activities or questions; expressions of (dis-) satisfaction</p> <p>Implicit and subsequent: purchasing orders, emails and messages expressing (dis-)satisfaction, follow-up questions and activities</p> <p>Explicit and (normally)</p>	<p>Explicit:</p> <p>Two raters use an analytic rating scale to score performance independently (ratings averaged); 4 categories scored on a scale of 1-5</p>

	subsequent: supervisors and managers assess the success of the event (presentation) by feedback/promotion/sanctioning	
--	---	--

7.4.2.2 Individual speaking task

In this chapter, the author utilizes the Task characteristics framework to analyze the characteristics of the individual speaking task (individual presentation/ short talk). The findings will be used in Chapter 7.4.3 to develop test task specifications for actual use in the educational domain (Phase 2 of the research). The following tables summarize the test task analysis, showing how the Task characteristics framework can be used as a Test task characteristics framework with the purpose of comparing TLU tasks to potential test tasks (see the Tables 7.15- 7.19 below).

7.4.2.2.1 The rubric

The rubric in the TLU situation is often implicit, with participants decoding contextual cues and responding to them. In testing contexts, the rubric is an interface through which the test taker collects and processes input and responds to the task prompt (see Table 7.20 below).

Table 7.20. Comparison of task characteristics of the TLU and test – individual speaking task (the rubric)

Characteristics	TLU situation	Test task
Rubric		
Specification of the objective	Implicit in TLU situations: to deliver a short talk expressing opinions about a certain problem/situation/issue to an interlocutor (or small audience)	Explicit: to deliver a short (monologue-like) talk on a given business topic (problem) after a short preparation time, and be ready to answer impromptu questions
Procedures for responding	Implicit: interact orally in English, explaining own point of view in a short monologue	Explicit: See the prompt, prepare notes (1min) and respond (1 min) orally in English, explaining own point of view in a short monologue
Structure of the communicative event		
Number of tasks	Varies by the number of questions asked/ one task	One task involving the preparation of an answer

	involving the preparation of an answer	
Time allotment	3-5 minutes	2 minutes
Evaluation		
Criteria for correctness	The interlocutor's satisfaction with provided argumentation/ problem or issue resolved/ the response addresses the issue in its entirety	Overall comprehensibility; interaction with the audience; overall impression; the problem/situation addressed in a clearly structured and organized talk with arguments/examples provided
Rating procedures	Implicit: embedded in the communicative event. The interlocutor responds to the talk, stating their satisfaction with the argumentation/explanation.	Explicit and formal: One rater using a holistic rating scale to score performance; Scores on a scale of 0-5

7.4.2.2.2 *The input*

The following table summarizes the comparison between the input used in the TLU situation for an individual speaking task to the more explicit rubric that must be developed in a speaking assessment. The characteristics pertaining to the input help test developers compare the characteristics of the TLU situation to those of the test setting, and attempt to replicate what they can in order to enhance situational and interactional authenticity of the assessment (see Table 7.21 below).

Table 7.21 Comparison of task characteristics of the TLU and test – individual speaking task (the input)

Characteristics	TLU situation	Test task
Input		
Prompt		
Features of the LSP context		
Setting	The settings vary from one context to another. Some of the shared characteristics are as follows: office/ business premises; table or a booth, could be office cubicle as well; printed material/ various objects/ computer screen, Internet connection.	Classroom, well lit and with the AC controlled temperature (heating in the winter season); 1 instructor desk facing 20 student desks (overall seating capacity of 40 students); the task takes place in the front of the room, with the instructor and a student sitting, facing each other; one long whiteboard, one overhead projector with the screen; the student (test-taker) writes notes on a piece of paper provided by the instructor, and then responds orally

Participants	One person in the role of the speaker, usually no more than 2-3 other people who listen to the talk. Usually unfamiliar to the speaker, people come to ask explanation/solution to the problem, seek advice or are curious about the description provided by the speaker.	One student at a time, and one instructor in the role of an interlocutor. Other students in the room, observing, not commenting.
Purpose	<p>Purposes vary from context to context. Some of the most recurring purposes include but are not limited to the following:</p> <ul style="list-style-type: none"> - to (individually) promote a product / service (in front of: business associates, prospective clients/ an existing client) - to (individually) give an overview of the company's mission, vision, range of products/service/ future plans for expansion/new markets (at fairs, exhibitions, joint presentations, Chamber of Commerce events, cluster events, etc.) - to demonstrate how a product works/ how a service is provided for prospective client(s) focusing on pros and cons and the company's relative standing in comparison to the competitors (price, quality, maintenance, warranty duration, extra services, etc.) - To address questions/ complaints/ and various issues that may arise. To defend a product/service/solution in front of consumers and other interested parties 	The purpose of the individual speaking task is to assess the spoken English ability of individuals to prepare and deliver a mini-presentation (up to 1 minute) on a business topic
Form and content	Time-limited language production provided by an individual in a question/answer format	Extended production by a test taker, presenting to one interlocutor.
Tone	Friendly and professional; restrained and fair; analytical	Friendly and professional; analytical
Language	World English with varying degrees of a foreign accent	English language (regardless of variety and accent)
Norms of interaction	Business representative/ colleague/client/interested parties	Business representative/ client; superior/inferior in business

	interaction. On the continuum from casual to formal	hierarchy; on the continuum from semi-formal to formal
Genre	Question/answer session; answer in the form of a short monologue-like presentation	Question/answer session; answer in the form of a short monologue-like presentation
Problem address to	Implicit in the TLU: <ul style="list-style-type: none"> - to address various issues/problems/complaints/ - to explain/describe/justify various aspects of a product/service/situation/activity 	Explicit: <ul style="list-style-type: none"> - to address various issues/problems/complaints/ - to explain/describe/justify various aspects of a product/service/situation/activity
Input data		
Format		
Visual	Written instructions/ directions; product specifications/ portfolio/ rules and procedures/ terms and conditions/ manuals	Written details pertaining to the task prompt
Audio	Questions from the interlocutors (clients, colleagues, business partners)	Spoken prompt (by the interlocutor/assessor); follow-up questions by the assessor
Vehicle of delivery	Live; oral	Live; oral
Length	1-2 hours to study the input data	1 minute for reading the prompt and prepare
Level of authenticity		
Situational	by definition	Shares some characteristics with potential TLU situations – moderate to limited situational authenticity
Interactional	Engaged (on a continuum from somewhat engaged to deeply engaged)	Deeply engaged

7.4.2.2.3 *The expected response*

Some of the characteristics of the expected response are deeply contextualized and embedded in a TLU situation. However, in a testing situation test takers need to know how to respond to the task prompt. The characteristics of the expected response help test developers analyze situational and interactional authenticity of the task, without compromising other qualities of a good testing practice. The following table summarizes the analysis between the characteristics of expected response in a TLU and in a test task (see Table 7.22 below).

Table 7.22. Comparison of task characteristics of the TLU and test – individual speaking task (the expected response)

Characteristics	TLU situation	Test task
Expected response		
Format		
Written	May include visuals and printed material	
Oral	Oral explanation of the problem, sometimes accompanied by demonstrating how something works	Oral explanation of the problem, accompanied with supporting arguments/examples
Physical	Manipulating a piece of equipment	
Type of response		
Selected		
Limited production		
Extended production	Extended response (5-10 min)	Extended but time limited production
Response content		
Nature of language	Vocabulary appropriate to the topic and audience	Vocabulary appropriate to the topic and audience
Background knowledge	Topic-related knowledge, economics, corporate culture norms	Topic-related knowledge, economics, corporate culture norms
Level of authenticity		
Situational	Building on the problem stated by the interlocutor, the presenter responds to the best of their knowledge. Situational authenticity is embedded in the context	Situational authenticity is somewhat limited to the test method; however the task shares some of the characteristics with a TLU task
Interactional	Deeply engaged	Deeply engaged

7.4.2.2.4 *The interaction between the input and expected response*

In a testing context, when performing on an individual speaking task where they are expected to produce a short monologue on a given topic, test takers rely on the input (prompt and input data) to provide them with processing materials based on which they will construct the response. The nature and quantity of input data vary from context to context, affecting the reactivity, scope and directness of the interaction between the input and expected response. The summary of the comparison of the two sets of characteristics is provided in Table 7.23.

Table 7.23. Comparison of task characteristics of the TLU and test – individual speaking task (the interaction between the input and response)

Characteristics	TLU situation	Test task
The interaction between the input and response		
Reactivity		
reciprocal: non-reciprocal	Highly reciprocal (all parties need to adapt as necessary so as to ensure mutual comprehension)	Somewhat reciprocal
Scope		
broad-narrow	Very broad	Narrow
Directness		
dependent upon input: dependent upon background knowledge	It can be anywhere on the continuum from indirect to fairly direct, depending on the following: <ul style="list-style-type: none"> a) the speaker can attend to the task by relying on their background/topical/technical knowledge, b) the speaker has to rely on the input (e.g. equipment manual) to respond. 	Fairly indirect: Test takers have to rely on their background knowledge and personal experience to respond to the prompt.

7.4.2.2.5 The assessment

The comparison between the characteristics of assessment in the TLU context and in the test context indicates the following (Table 7.24):

1. An individual speaking task in the TLU may refer to addressing various language functions and speaking purposes. The TLU construct definition summarizes the abilities that language users need to demonstrate in order to address those purposes. Due to the constraints of practicality and validity, not all components of the TLU construct can be translated into the test task construct. In a texting context, the construct definition is less comprehensive, aimed at targeting a more narrowly identified construct of speaking ability.
2. The criteria for correctness and rating procedures are predominantly implicit in the TLU situation, indicating that the TLU tasks employ the principles of the indigenous assessment to rate the speaking performance. The assessment criteria and rating procedures in assessing the speaking test task are explicit, helping test

developers/assessors rate the performance by maximizing the qualities of reliability and validity throughout the process.

Table 7.24. Comparison of task characteristics of the TLU and test – individual speaking task (the assessment)

Characteristics	TLU situation	Test task
Assessment		
Construct definition	Specific purpose language ability is quite complex in the TLU situations in the observed TLU domain. Some of the shared characteristics are as follows: general business terminology, customer relations (including customer support) terms, the knowledge of marketing and pan-technical terminology; the use of declaratives, tag questions and rhetorical questions, indirect and Wh-questions; the cohesive use of discourse markers; organization knowledge of process structure, transitions and turn-taking strategies; use of heuristic, ideational, and manipulative functions; common idioms and cultural references; problem-solving skills; strategic use of expository and conversational styles, using comprehension checks. Background knowledge: ability to elaborate on the topic by using the terminology the client/interlocutor can easily understand; awareness of service provider/client business conventions	The task sets out to assess the following: overall English language comprehensibility; the use of Business English vocabulary covered by the course; coherency and use of discourse markers in speech; spoken grammar, fluency and pronunciation; interaction with the interlocutor(s).
Criteria for correctness	Indigenous criteria: presentation skills, pronunciation and comprehensibility, voice projection, cultural awareness, content/background knowledge, presenters' personality and experience (friendly, professional, responsive, learned)	Explicit criteria: The assessor rates the performance by checking for structure and clarity of ideas, coherence, pronunciation, spoken grammar, and vocabulary; norms of politeness (since it is relevant to the TLU domain)
Rating procedures	Implicit: the interlocutor(s) may make private judgments of the communicative act, or they can take a more active approach and	Explicit: the assessor rates the performance rated by using a holistic rating

	ask questions, make/withdraw the purchase/order; expressions of (dis-)satisfaction Explicit: supervisors and managers assess the success of the communicative event by feedback/promotion/sanctioning	scale on a band of 1 to 5
--	--	---------------------------

7.4.3 Test task specifications

Building on the works of Bachman and Palmer (1996) and Douglas (2000), the author will present the test task specifications that ensue from the analysis above. According to Bachman and Palmer, test tasks are derived from TLU task types, and then modified in the process of the test development so as to meet the criteria for test usefulness (1996). Building on the idea of test usefulness, Douglas (2000) states that “in making the transition from the analysis of the target language use tasks to test tasks [...] TLU tasks are either adapted or eliminated altogether” (p.115). The objective of the TLU task analysis in this study is to produce two task types:

- 1) that are relevant to the majority of the TLU situations analyzed,
- 2) that are easy to adapt to test tasks without interfering much with qualities of test usefulness (in particular the quality of authenticity), and
- 3) that are in line with learning objectives in the setting for which they are intended (English language course at the Faculty of Economics, University of Kragujevac).

The TLU analysis outlines two recurring tasks on whose relevance to the TLU situations the majority of subject specialist informants agree: 1) a group oral presentation, and 2) a short individual presentation. Both tasks, however, require extended spoken production in Bachman’s sense; and to help test developers distinguish between the characteristics of these two test tasks the author will provide task specifications to ensure the complete coverage of the construct and attend to the qualities of test usefulness.

Building on the test task specifications model proposed by Bachman and Palmer (see 5.3.2), test task specifications will be designed and will include the following:

- the purpose of the test task,
- the definition of the construct to be measured
- the learning outcomes addressed by the construct definition
- the characteristics of the setting of the test task,
- time allotment,
- instructions for responding to the task,
- scoring method,
- plan for evaluating test usefulness qualities.

It should be noted that the author has made three changes to the model of test task specifications proposed by Bachman and Palmer:

- 1) Building on the recommendations made by Green (2014) that task and items specifications can be created as learning outcomes or as a means of capturing the features of real life language, combines both approaches to bridge the gap between two domains. Consequently, the author includes the identification of the learning outcomes addressed by the construct definition here, because these two types of tasks are relevant to both domains: real life and the domain of education. The latter, being the setting for the task delivery, operates in terms of learning outcomes that any assessment should reflect,
- 2) The author excludes the characteristics of input, response, and relationship between input and response, as they are embedded in the task design and instructions provided to test takers,
- 3) Following Douglas's model, the author includes the plan for evaluating test usefulness qualities (or the qualities of good testing practice, as Douglas name them, 2000:118).

7.4.3.1 Group speaking task – test task specifications

7.4.3.1.1 The purpose

The purpose of the group speaking task is to assess the ability to deliver a group oral presentation on a business topic in English language. The task is a part of formative assessment plan.

7.4.3.1.2 Construct definition

The task sets out to assess the following: overall English language comprehensibility; genre-appropriate and topic-appropriate vocabulary; marketing-related vocabulary employed, spoken grammar, fluency and pronunciation; communication skills and confidence; appropriate non-verbal communication; use of transitions, persuasive language; use of discourse markers; use of visuals; interaction with the audience.

7.4.3.1.3 Learning outcomes

The construct is in line with the following learning outcomes outlined in the course syllabus:

Students will be able to:

- deal with less routine situations and explain why something is a problem
- exchange, check and confirm information
- give or seek personal views and opinions in a discussion
- seek and report other people's views and opinions
- give descriptions
- prepare and deliver oral presentations
- deal with less routine situations and explain why something is a problem
- deliver presentations of an informative nature
- structure a presentation into its component parts and use transitions to move from one point to another
- use appropriate body language (gestures, facial expressions, eye-contact, and posture) and oral communication (voice projection, fluency, pronunciation, intonation) to convey a message
- demonstrate field-specific background knowledge (*Marketing*)
- use suitable visual aids to enhance their presentation and reinforce their message
- interpret visuals (graphic organizers – tables, charts, graphs; interpret visual/aural input – video/audio recordings)
- engage audience when presenting
- use technical, field-specific vocabulary accurately and effectively

7.4.3.1.4 *The characteristics of the setting of the test task*

Physical setting

The group presentation task takes place at a theater with a stage and podium. There are abundant resources for displaying visuals to help the audience get a full picture of the topic presented on: blackboard, flip-chart and flip-chart holder, an overhead projector connected to a desktop computer (with the Internet connection) and a large screen above the blackboard, a presentation laser pointer/PowerPoint clicker. The presenters have an option to choose whether they want to use lapel and/or hand microphones (wireless). The audience is seated in the rows of seats and student desks surrounding the stage and the podium from three sides. The room is well lit and the room temperature is controlled by the AC. The theater is a large room with the seating capacity of 300.

Participants

In each session there are 25 students in the audience including the presenting group. Male and female students are aged between 20 and 30. There are two instructors assessing the performance; they are sitting unobtrusively behind other students who assume the role of the audience. The presenters are familiar with the audience and instructors.

Time allotment

This kind of task requires thorough preparation and delivery. For this reason, the time allotment is divided into two phases: 1) preparation, and 2) delivery.

Phase 1 – Task preparation

The preparation phase was implemented during the spring semester as per the following timeline (Table 7.25):

Table 7.25 Timeline for the presentation task

Week 6: task announcement, group assignment, topic proposal
Week 7: rating scales presented to students (self-assessment, peer-assessment, and instructor-assessment)
Week 8: students start research (data collection, online research work)
Weeks 9-11: students receive feedback from their instructors

Phase 2 – Presentation delivery

The delivery phase takes two weeks to implement. Considering that the number of students who took the task was 150, they are divided into 6 time slots attended by 25 students. The presentations take place on different days, starting at 10am.

Presentation delivery time: 5 minutes to set up, and up to 15 to present, depending on the number of students in a group. Since students are divided into groups of 3 and 4, the former present up to 10 minutes, and the latter up to 15.

Weeks 12 – 14: groups present in front of two assessors and audience in the theatre or seminar rooms (the location and timings announced on the notice board and online, on the Faculty of Economics homepage)

Instructions for responding to the task

The following set of instructions is provided to test takers (presenting groups):

“In groups of three to four, students will deliver an oral presentation on a company/product/service of their own choice. They will choose among companies conducting business operations on the territory of the Republic of Serbia (the company itself can perform business internationally or locally). Students are required to demonstrate their background knowledge related to the field of marketing, and talk about the topic assuming the role of a product/service/company promotion team. Each group will conduct a research (on the Internet and/or live, preparing surveys and questionnaires related to the topic of their choice), and design a PowerPoint presentation, including audio/video recordings and graphic organizers (charts, graphs, tables) ensuring that selected visuals support the collected data. The group presentation structure should incorporate both verbal-and non-verbal communication skills, following the principles of effective presentation delivery (Laws, 2010; Powel, 2011, AUM, 2016). The presentation should be maximum 10 minutes long for the groups of three/ 15 minutes long for the groups of 4 students. Students will distribute the roles ensuring that each participant is assigned equal time for delivering their part of the presentation.”

7.4.3.1.5 Scoring method

The task is scored objectively by two trained instructors who teach students and who are familiar with them. To ensure objectivity of scoring, the assessment criteria are assessed by one analytic and one holistic rating scale, an analytic and holistic (two rating scales are provided for the research purposes, to test hypotheses H1 and H2, see 6.2.1 above).

The analytic rating scale includes the following criteria (see Appendix J):

- Group dynamics and Presentation structure
- Visuals and Audience engagement
- Non-verbal communication
- Verbal communication
- Grammar and vocabulary

The criteria above are assessed on a scale of 1 to 5, with descriptors of performance assigned to each point (1 – poor, 2 – below expectations, 3 – meets expectations, 4 – above expectations). The task is assigned 10 points total, requiring that assessors multiply the points in the scale by 2 in order to get the total score.

The holistic rating scale includes a combination of the criteria mentioned above, and scores the performance on the scale from 1 to 10, with 1-2 assigned to a response which is *unsatisfactory*, 3-4 to a *poor* performance, 5-6 to a performance which is *below expectations*, 7-8 to a performance that *meets expectations*, and 9-10 to a performance that is *above expectations* (see Appendix K). The task is assigned 10 points total, with the number of points in the scale matching the total number of points assigned to the task.

7.4.3.1.6 Plan for evaluating test usefulness qualities

- 1) Reliability: inter-rater reliability is achieved through standardization sessions, trialing and piloting the rating scales.
- 2) Validity: test-takers' self-assessment of language ability at the beginning and the end of the semester by means of self-evaluation "Can-do" questionnaire. The data obtained in this manner are corroborated with final oral exam results and placement test results. Finally, students assess the validity of the task and assessment process by means of a questionnaire investigating their attitudes about authentic tasks and forms of assessment.

- 3) Situational authenticity: test developers will compare the characteristics of actual test task administration to the data obtained from subject specialist informants.
- 4) Interactional authenticity: the end of semester questionnaire will be given to students to investigate their involvement in the task.
- 5) Impact/Consequences: interview instructors about how the task affects student learning and final grades
- 6) Practicality: the venue and equipment provided by the Faculty of Economics; invigilators not required as the task is delivered orally; two raters are the instructors teaching the course. Time is the main constraint: it takes approximately 20 minutes for one group to set up and deliver the presentation; another 4-7 minutes to fill out the rating scale and add comments; finally, it takes a week to check the rating scales and announce the grades.

7.4.3.2 Individual speaking task – test task specifications

7.4.3.2.1 The purpose

The purpose of the individual speaking task is to assess the spoken English ability of individuals to prepare and deliver a mini-presentation (up to 1 minute) on a business topic. The task is a part of formative assessment plan.

7.4.3.2.2 Construct definition

The task sets out to assess the following: overall English language comprehensibility; the use of Business English vocabulary covered by the course; coherency and use of discourse markers in speech; spoken grammar, fluency and pronunciation; interaction with the interlocutor(s).

7.4.3.2.3 Learning objectives

The construct is in line with the following learning objectives outlined in the course syllabus:

Students will be able to:

- deal with less routine situations and explain why something is a problem
- exchange, check and confirm information
- give or seek personal views and opinions in a discussion

- give descriptions
- deliver short oral presentations after short preparation
- deal with impromptu questions
- deliver presentations of an informative nature
- structure a presentation into its component parts and use transitions to move from one point to another
- use appropriate body language (gestures, facial expressions, eye-contact, and posture) and oral communication (voice projection, fluency, pronunciation, intonation) to convey a message
- demonstrate the ability to apply the norms of politeness
- use technical, field-specific vocabulary accurately and effectively

7.4.3.2.4 The characteristics of the setting of the test task

Physical setting

The task takes place in a classroom that is well lit and with the AC-controlled temperature (heating in the winter season). There is one instructor desk facing 20 student desks (overall seating capacity of 40) in the room. Apart from the chairs and desks, there is one long whiteboard, and overhead projector and the screen. The task takes place in the front of the room, with the instructor and a student sitting, facing each other. The student (test-taker) writes notes on a piece of paper provided by the instructor, and then responds orally.

Participants

There is one student at a time, and one instructor in the role of an interlocutor. Other students are in the room, observing, not commenting.

Time allotment

This kind of task requires short preparation and delivery. For this reason, the time allotment is divided into two phases: 1) preparation (1 minute), and 2) delivery (1 minute). The overall time allotted for this task is 2 minutes per test taker.

Instructions for responding to the task

The task and instructions for responding are based on the CUP's Business English Certificate Speaking section of the exam. More specifically, this speaking task has a lot in common to Part Two of the Speaking section of the BEC exam (Cambridge University Press, 2002; 2004; 2009).

The following set of instructions is provided to test takers (presenting groups):

“In this task, you are asked to give a short talk on a business topic. You have to choose one of the two topics provided on the paper in front of you, and then talk for about ONE minute. You have ONE minute to prepare your ideas (you may use the blank notepaper provided on the desk in front of you).

Think about the topic and support your arguments by examples. Your response will be assessed for structure and clarity of ideas, coherence, pronunciation, spoken grammar, and vocabulary.”

Table 7.26 Example of a short individual presentation task

SHORT TALK ON A BUSINESS TOPIC	(2 min) _____/5pts.
In this task, you are asked to give a short talk on a business topic. You have to choose one of the two topics provided on the paper in front of you, and then talk for about ONE minute. You have ONE minute to prepare your ideas (you may use the blank notepaper provided on the desk in front of you.	
TOPICS:	
<i>A: What is important when...?</i>	
<u>Deciding the price of a product</u>	
- cost of production	
- cost of similar product	
- the size of the market	
<i>B: What is important when...?</i>	
<u>Arranging a social event for clients</u>	
- types of activities	
- cost of event (food, drink, entertainment)	
- venue	
Think about the topic and support your arguments by examples. Your response will be assessed for structure and clarity of ideas, coherence, pronunciation, spoken grammar, and vocabulary.	
<i>Based on BEC Speaking tasks (Cambridge University Press, 2002; 2004; 2009)</i>	

7.4.3.2.4 Scoring method

The task is scored objectively by a single instructor (who may but does not have to be teaching the students taking the exam). To ensure the objectivity of scoring, the assessment criteria are assessed by a holistic scoring scale with the points in the range of 0 – 5. The task is assigned 10 points total, requiring that assessors multiply the points in the scale by 2 in order to get the total score. The following criteria are included in the holistic scale: structure and clarity of ideas, coherence, pronunciation, spoken grammar, vocabulary, and norms of politeness. Given the fact that scoring criteria are derived directly from the TLU domain, politeness plays an important role in interpersonal communication.

The following is an example of a top-scoring performance, based on the holistic scale developed specifically for the short individual presentation task (Table 7.27; for the full scale, see Appendix M).

Table 7.27 Example of a holistic scale (score 5) to rate performance on the individual speaking task

5	Excellent	Can communicate ideas clearly and in a structured manner, providing appropriate examples; uses discourse markers and speaks coherently; pronunciation clear; minor grammar mistakes; topic-appropriate Business English vocabulary
---	-----------	--

7.4.3.2.5 Plan for evaluating test usefulness qualities

- 1) Reliability: intra-reliability enhanced by period rater “refresher” training sessions.
- 2) Validity: test-takers’ self-assessment of language ability at the beginning and the end of the semester by means of self-evaluation “Can-do” questionnaire. The data obtained in this manner are corroborated with final oral exam results and placement test results. Finally, students assess the validity of the task and assessment process by means of a questionnaire investigating their attitudes to the whole process.
- 3) Situational authenticity: test developers will compare the characteristics of actual test task administration to the data obtained from subject specialist informants.
- 4) Interactional authenticity: the end of semester questionnaire will be given to students to investigate their involvement in the task.

- 5) Impact/Consequences: interview instructors about how the task affects student learning and final grades
- 6) Practicality: the venue and equipment provided by the Faculty of Economics; invigilators not required as the task is delivered orally; one rater, who may be one of the instructors teaching the course. Time is the main constraint: although it does not take long for one candidate to prepare a short talk and deliver it, there are more than other 20 test takers waiting for their turn, which may cause the fatigue both in the instructor and students (potentially affecting the quality of reliability).

8 Phase 2

8.1 Participants in Phase 2

The research includes data collected from students enrolled in the course *English Language 2*, in the academic 2016/2017, on condition that they had successfully completed the course *English Language 1*. Bearing in mind that around 500 students enrol in business modules each year, the representative sample of 150 students (30%) participated in the research (upon signing the consent form, Appendices D and E).

Students who enrol in business courses choose between two study programs - *Business Economics and Management* and *Economics*. Each of the study programs is further divided into following modules (Table 8.1):

Table 8.1 Study programs at the Faculty of Economics, University of Kragujevac (Academic year:2016/2017)

<i>Business Economics and Management</i>	<i>Economics</i>
<i>Accounting and Business Finance</i>	<i>General Economics</i>
<i>Marketing</i>	<i>Finance</i>
<i>Management</i>	<i>Stock Exchange</i>
<i>Tourism and Hospitality</i>	<i>Banking</i>

The participants in the research include the students enrolled in the following modules: *Management*, *Accounting and Business Finance*, and *Management*. The selection above is made for the following two reasons:

- students enrolled in these modules follow the same English language syllabus, whereas all other modules have different learning objectives and different number of contact hours, and
- students enrolled in these modules share the majority of mandatory and optional courses, including the course in *Marketing* (mandatory for students enrolled in *Marketing* module, but elective for the other two modules).

The students who agree to participate in the research have to meet additional requirements:

- students have successfully passed their exam in *English Language 1*;
- students have completed a minimum of 8 years studying English at primary, secondary and tertiary levels of education;
- students have attended the course in *Marketing*, in the first semester of their second year of university studies (the course is, however, mandatory for the *Marketing* module)
- students have signed the consent form and agreed to take the placement English test.

All students took the Placement test (See chapter 8.2.1 below) at the beginning of the semester. As per the research proposal, the groups were formed in such manner that the average point per group is the same (30 out of 60 points on the Placement test). According to the placement test key, this number of points corresponds to B1 level of the Common European Framework of Reference (Appendix C).

Based on their respective module and placement test results, student participants were divided into two experimental and one Control group as per the table below:

Table 8.2 Groups (student participants representing the domain of education)

Group	Module	Placement test result (averaged - per group)	CEFR Level (averaged - per group)
Group 1	<i>Management</i>	30	B1
Group 2	<i>Accounting and Business Finance</i>	30	B1
Group 3	<i>Marketing</i>	30	B1

8.2 Research instruments

8.2.1 Placement test

At the beginning of semester 2, as well as at its end, students take parallel versions of a *pen and paper* (P&P) placement test of English produced by Oxford University Press and

Cambridge ESOL. The intended purpose of this Quick Placement Test (Appendices A and B) is to provide instructors with a fast and reliable measure of test takers' proficiency in English, with results aligned with ALTE and CEFR levels of proficiency. Test scores can be used for placement decisions and for grouping students according to their respective levels. Considering the fact that parallel versions of this test are administered, they share the same task characteristics and target the same constructs of language proficiency. In the text below, Versions 1 and 2 will be discussed as a single version.

The pen and paper Quick Placement Test consists of two parts assessing test takers' *reading* skills as well as the knowledge of grammar and vocabulary tested jointly as *structures*. Part One includes items 1 to 40, with questions aimed at students who are at intermediate level or below. Part Two includes items from 41 to 60. These items contain questions that are more difficult and they cover all ALTE levels (consequently, all of the CEFR levels, as well), including C2. The test is intended for test takers of all levels and all ages, and can be administered either as a pen and paper test, or as a computer-based test (both versions of the test are available, but the choice depends on the qualities related to test practicality, such as equipment, facilities, invigilators, test administrators, etc.). It should also be noted that computer-based version of the test and pen-and-paper version do not test exactly the same constructs, given that listening skills are tested only in the computer-based version of the test. According to Geranpayeh (2003), the Quick Placement Test can be used in the following ways: (1) before the course starts, so that administrators can use the scores to make student placement decisions based on the scores; (2) on the first day of the course, since test scoring is fast and reliable; (3) during the course, to place students who enroll late for some reason; (4) at any time to decide whether students are eligible for particular courses (p.8).

All three groups of students take a placement test of English at the beginning of semester for the following reasons:

- to provide the researcher with general insight into students' strengths and weaknesses;
- to determine the students' proficiency levels;
- to help the researcher divide students into three experimental groups (including one Control group), with the same average (as well as CEFR level) per group;

- to provide the basis for comparison with the exit test results (a parallel version of the Placement test, administered at the end of the semester, and used as a placement test for the following semester);
- to corroborate the results of self-evaluation in order to validate H5 (see Chapter 6.2).

8.2.1.1 Placement test – participants

Students enrolled in the second year of studies at the Faculty of Economics, majoring in *Management, Accounting and Business Finance*, and *Marketing* volunteered to participate in the research. The number of students who volunteered to participate in the research was 272, out of which 179 were female and 107 male, aged between 20 and 34. Considering the research objectives, the participants' gender and age play no role in the study and will not be discussed any further.

8.2.1.2 Placement test structure and time allotment

There are 60 multiple-choice items testing Reading and Structures (grammar and vocabulary). Test takers are given 30 minutes to complete the test.

8.2.1.3 Task types

1) Tasks type 1

Reading tasks 1-5 are intended at assessing low-order reading skills (identification and basic metacognition in Alderson's sense (2000)), informational meaning (in Cohen's sense (1994)) as well as strategic competence and background (general knowledge) in a multiple-choice format with three options (A,B, and C) and one correct answer. The correct answer is assigned one point. The scoring is dichotomous: correct/incorrect (Cambridge ESOL, 2002, Example 8.1 below).

Example 8.1 Quick Placement test task assessing low-order reading skills

Questions 1 – 5

- Where can you see these notices?
- For questions **1** to **5**, mark **one** letter **A**, **B** or **C** on your Answer Sheet.

1

**Please leave your
room key at Reception.**

A in a shop
B in a hotel
C in a taxi

2) *Tasks type 2*

In accordance with the principles of communicative language testing, there are another five reading sections assessing grammar and vocabulary in context (Rea-Dickins, in Purpura, 2004). Indirectly they also target test takers' higher-order reading skills (for example, recognizing coherence in a text); however, the majority of items are specifically written so as to assess grammar and vocabulary in use (Cambridge ESOL, 2002, Example 8.2.)

Example 8.2 Quick placement test task assessing reading and structures

Questions 6 – 10

- In this section you must choose the word which best fits each space in the text below.
- For questions **6** to **10**, mark **one** letter **A**, **B** or **C** on your Answer Sheet.

Scotland

Scotland is the north part of the island of Great Britain. The Atlantic Ocean is on the west and the North Sea on the east. Some people (6)..... Scotland speak a different language called Gaelic. There are (7) five million people in Scotland, and Edinburgh is (8) most famous city.

Scotland has many mountains; the highest one is called 'Ben Nevis'. In the south of Scotland, there are a lot of sheep. A long time ago, there (9) many forests, but now there are only a (10)..... Scotland is only a small country, but it is quite beautiful.

6 A on **B** in **C** at

3) *Tasks type 3*

Multiple-choice questions within a sentence stem, targeting grammar structures and vocabulary (Examples 8.3 and 8.4). Items in the second part of the test (items 41-60) are considerably more difficult as they target test takers at a higher level of proficiency.

Example 8.3 Quick placement test assessing grammar and vocabulary

Questions 21 – 40

In this section you must choose the word or phrase which best completes each sentence.
 For questions 21 to 40, mark **one** letter **A**, **B**, **C** or **D** on your Answer Sheet.

21 The teacher encouraged her studentsto an English pen-friend.
A should write **B write** C wrote **D to write**

22 They spent a lot of time at the pictures in the museum.
A looking **B for looking** C to look **D to looking**

Example 8.4 Quick placement test assessing higher level proficiency grammar and vocabulary

58 A lot of the views put forward in the documentary were open to
A enquiry **B query** **C question** **D wonder**

59 The new college..... for the needs of students with a variety of learning backgrounds.
A deals **B supplies** **C furnishes** **D caters**

60 I find the times of English meals very strange – I'm not used dinner at 6pm.
A to have **B to having** **C having** **D have**

8.2.1.4 *Quick Placement Test administration*

The test was administered at the very beginning of the course *English language 2*, in Week 2, when student enrolment was finalized. The number of students who took the exam was 272. It took a week for two instructors to grade the test, and announce the results, based on which the instructors selected 150 students enrolled in three modules to participate in the research (following the students' signing the consent form to participate). A parallel version of

the same test was administered to the same 150 participating students at the end of the course. It should be noted that was not in line with the intended use of the test, since it is not developed for use as a end-of-the-course test of English language. However, its administrations can be justified for two reasons:

- 1) to secure student participation, and
- 2) to make placement decisions for the following course (English language 3).

8.2.1.5 Quick placement test – the analysis of test usefulness qualities

Quick Placement Test was chosen for the following reasons related to the qualities of test usefulness: practicality, reliability, and validity.

8.2.1.5.1 Practicality

Before the test is administered, test booklets need to be printed out. The test booklet contains 10 pages, including the cover page. To ensure the test security, two English instructors printed out and copied the test booklets in the Exam Control room, and stored the booklets safely until the test administration day. As per instructions, test administration time is 30 minutes, with additional 30 minutes required to check student IDs and arrange seating before the start. The venue – the theatre that seats 300 students - and 8 test invigilators were provided by the Faculty of Economics, with two English language instructors with the floating duty (helping students in the case of difficulties with test administration). Invigilators and floaters performed their duties within regular work hours. Apart from the printing and copying expenses, borne by the Faculty, test administration did not require any other financial resources.

8.2.1.5.2 Reliability

Quick Placement test employs a multiple-choice format to test reading, grammar and vocabulary, based on the Communicative model of language proficiency. Its scoring is facilitated by the use of transparency overlay sheets over the answer sheet enabling fast and reliable scoring process. In order to secure the reliability of rating process, raters performed a peer-marking check, by randomly selecting 10 papers to check if the scores were arrived at accurately. Scores on the pen and paper test are reported on a scale out of 60, with points corresponding to ALTE and CEFR levels. At the same time, the scores can be used as an indicator of the level in terms of

BEC (Business English Certificate), which is of relevance for the research conducted with business majors(see Table 8.3).

Table 8.3 Test scores aligned to ALTE, CEFR and BEC frameworks

Points	ALTE		CEFR	Relevance to Business English levels
	Level	Description		
0-10	0.1	Beginner		
11-17	0.2	Breakthrough	A 1	
18-29	1	Elementary	A2	
30-39	2	Lower-Intermediate	B1	
40-47	3	Upper – Intermediate	B 2	BEC P Business English Certificate Preliminary
48-54	4	Advanced	C1	BEC V Business English Certificate Vantage
55-60		Very Advanced	C 2	BEC H Business English Certificate Higher

8.2.1.5.3 Validity

Ardeshir Geranpayeh, a member of the English Quick Placement Test validation team, states that the test itself was validated in three phases. First, the test format was validated in an international validation project started with the idea to determine content and construct validity with respect to accuracy at placing test takers at appropriate proficiency levels. Second, the score consistency was validated and proved to be reliable in two successive administrations. Three, there was the final stage of determining the equivalence between the two test modes: pen and paper and computer-based test (2002:9).

8.2.1.5.4 Authenticity

The authenticity of the Quick Placement Test is relatively low, both in terms of situational and interactional authenticity. However, considering the fact that authenticity of an assessment often depends on the purpose of the assessment, it should be noted that the test was

intended to be used as a placement tool, and considering that it meets other requirements of test usefulness, it was chosen as a reliable indicator of test takers proficiency level.

8.2.2 Task-based approach – authentic speaking tasks

Upon rating the performance on the Placement test, at the beginning of the semester, the researcher will divide student participants into three groups (as per the module and Placement test results, see 8.1 above) sharing the same group average. The participants in Groups 1 and 2 will be exposed to authentic speaking tasks (following the task-based approach to assessment, see Chapter 3.1.3 above) developed during Phase 1 of the research - a group presentation task and an individual presentation task (see Chapter 7.4.3 for Test task specifications). The tasks were developed in collaboration with subject specialist informants, involving the researcher in the role of a test developer. The Phase 1 deliverables – authentic speaking tasks- share the characteristics of the TLU speaking tasks identified in the real life domain, based on the comparison in the Test task characteristics framework. Throughout the semester, student participants will engage in authentic speaking tasks, in a series of formative assessment sessions. The exposure to authentic speaking tasks involves critical elements pertaining to authentic assessments (see Chapter 4.3):

- Challenge
- Outcome: performance or product
- Transfer of knowledge
- Metacognition
- Accuracy
- Environment and tools
- Feedback
- Collaboration

The third group of students, enrolled in *Marketing* module, will take the role of a Control group where students are exposed to tasks developed according to course syllabus requirements. The assumption is that test tasks developed in line with this principle possess a lower level of situational and interactional authenticity than those created within a task-based approach.

8.2.2.1 Differences among experimental groups- task format

As outlined above, Groups 1 and 2 will respond to test tasks corresponding to specific purpose speaking tasks identified within the real life domain:

- a group speaking task – presentation (see 7.4.3.1 above), and
- an individual speaking task – short talk/mini-presentation (see 7.4.3.2 above).

The authentic speaking tasks developed during Phase 1 are to be applied in Phase 2 for formative purposes. Student participants will receive no summative assessment grades that might affect their overall course score. The following table outlines the differences among task formats the groups will be exposed to:

Table 8.3 Task format differences (per group)

	Group 1	Group 2	Control group
Group speaking tasks (presentation)	√		
Individual speaking tasks (a short talk/ presentation)	√	√	
Structured speaking tasks (structured Q&A format)	√	√	√

Subjects in Group 1 are students enrolled in *Management* module. They will be expected to demonstrate their ability to participate in the following task types: group speaking tasks, individual speaking tasks, and structured speaking tasks. Throughout the semester, they will be exposed to mini-tasks aimed at enhancing their group presentation skills (group dynamics and transitions), individual presentation skills, collaboration skills, research skills, and delivery skills.

Subjects in Group 2 are students enrolled in *Accounting and Business Finance* module. They will be expected to demonstrate their ability to engage in the following speaking tasks: individual presentation tasks and structured speaking tasks. Throughout the semester, they will be exposed to mini-tasks aimed at improving their self-initiativeness, individual presentation skills, research skills and delivery skills.

It should be noted that Groups 1 and 2 will receive the instructor's detailed feedback upon every class presentation. The feedback will include both positive and negative aspects of the performance on the task, tackling evaluation criteria and possible corrective measures that students can take in order to enhance their presentation or speaking skills.

Unlike the first two groups, subjects in the Control group will be exposed to the syllabus-based tasks which require that test takers respond to highly structured tasks that foster the development of micro- rather than macro-skills.

8.2.2.2 Differences among experimental groups - evaluation criteria and feedback

The following difference in the approach to testing students' speaking skills is the one referring to their familiarity with evaluation criteria and the feedback they get from instructors. In line with authentic and formative testing requirements, the extent to which examinees are familiar with standards by which their performance is judged can help them improve over time. To this end, students in the first and second group will be familiarized with evaluation criteria administered by the means of holistic and analytic rating scales (see Chapter 5.3.3.1 above); they will receive a comprehensive feedback regarding their overall performance; they will be trained in rating their own as well as peers' performance; and, finally, they will learn how to monitor their own progress through the process of self-evaluation (by using the CEFR-aligned "Can-do" checklists).

From the beginning of the semester, participants in Group 1 will learn how to use analytic scales to evaluate their own and the performance of their peers. The Group 2 participants will be engaged in the same activity, but students in this group will be using holistic scales to rate their own and the performance of their peers.

The Control group, on the other hand, will receive instructors' feedback, based on holistic rating scales helping students get insight in their strengths and weaknesses. However, students in this group will not undergo a thorough semester-long training on rating their own or their peers' performance, and neither will they be asked to monitor their own progress based on a set of pre-determined criteria. Getting students familiarized with the use of rating scales is deemed relevant to the real life domain where constructive criticism and self-criticism at a work place are not only encouraged but also required by employers and other stakeholders. Differences related to the participants' exposure to evaluation criteria will help the researcher validate hypotheses H1, H3, and H4 (see Chapter 6.2.1 above).

8.2.3 "Can-do" checklists (survey)

Apart from data originating from formative and summative assessment conducted throughout the semester, the research will benefit from another set of data provided by student respondents (as well as Phase 1 participants – subject specialist informants). More specifically, a set of closed-ended checklists will be provided to both groups of participants, helping them rate and monitor their own knowledge and progress in terms of spoken interaction/production in English (Appendix O, parts A and B, and Appendix O, parts A and B). The same set of checklists is provided to subject specialist informants (see Appendix Q, parts A and B, and Appendix R, parts A and B), during Phase 1, with the intention of examining the desired speaking interaction/production English language skills (or more specifically, the CEFR level reflecting those skills) in their particular work settings. Subject specialist informants take the survey during Phase 1 of data collection, and their answers are statistically analyzed with those provided by student respondents. Both sets of responses are analyzed and compared at the end of Phase 2.

8.2.3.1 What are "Can-do" checklists?

Analyzing the possible role of CEFR descriptors in rating scale design (see Chapter 5.3.3.3 above), we identified the following potential uses of the Framework:

- to state the criteria to determining the attainment of learning objectives, and
- to describe different levels of proficiency (COE, 2000).

Building on this approach to using the CEFR scales, The Council of Europe has been supporting various projects aimed at utilizing the CEFR and expanding its lists of descriptors for languages. One of the deliverables funded by COE is a list of *Generic checklists for use in ELPs designed for language learners aged 15+*. Checklists containing descriptors or a set of statements starting with “I can” are provided by the Council of Europe as a “detailed inventory of communicative activity that can be used for regular goal-setting and self-assessment” (2015:1). The original checklists contain descriptors aligned with the CEFR levels, from A1 to C2, and cover both receptive and productive skills. They are intended for European Language Portfolio developers who work with language learners older than 15 years. These are the main reasons why the checklists are adopted as a self-evaluation instrument necessary for this research. The original checklists can be regarded as a starting point with new descriptors added, or the existing descriptors modified to suit a particular communicative event or a purpose. The author and the readers should be aware of the following guidelines, outlined by the authors of the checklists:

- it is not possible to create a comprehensive checklist that will encompass the full range of communication related to any CEFR level or activity, so new descriptors should be added over time;
- the more descriptors the checklist contains, the more effective it is in helping learners set and monitor their language learning goals;
- it has been suggested that when learners can perform at least 80% of the tasks/activities specified for a particular level and activity, it can be assumed that they have attained the level/activity in question;
- the checklists can be presented to learners in the form that supports monitoring and planning, so that they can set targets and monitor their own progress.

Adopting the checklists for self-evaluation and goal-setting purposes, the author of the thesis made the following amendments prior to conducting the surveys in which the checklists were used:

- although the checklists contain descriptors for all language skills, only spoken (production and interaction) skills are applied as survey instruments;

- descriptors were modified so as to fit the context of English for specific purposes in the context of business studies;
- descriptors used in the checklists provided to student respondents follow the “I-can” format, whereas the descriptors used in the checklists provided to subject specialist informants follow the “can-do” format. The reason for this lies in different purposes of the checklists. The former are used for students- self-evaluation and goal setting; they are supposed to indicate the progress that students have achieved during the course of their studies. The latter indicate the desired CEFR level in work settings, and for this reason the descriptors refer not to the respondents but to their prospective employees, hence the third person plural in the descriptors;
- descriptors are shuffled and arranged within a closed-ended survey, and then provided to respondents as a checklist to which they are supposed to respond by selecting the descriptor that is true to them. The process of shuffling is a necessary step to increase the objectivity of selections, as it seems fairly obvious that subsequent descriptors reflect the ability at higher CEFR levels, and in that case student respondents may select them intentionally to ‘inflate’ the actual levels, knowing that their language instructor may see the results. An additional objectivity measure was taken by the requirement that students should submit their checklist questionnaire as anonymous within their respective experimental groups;
- the author added the following set of instructions:
 - a) to subject specialist informants: “Read the descriptors below and tick (√) ONLY THE BOX showing what you think your prospective employees should know how to do“ (see Appendices, Q and R);
 - b) to student respondents in Experimental groups 1 and 2: „Read the descriptors below and tick (√) ONLY THE BOX showing what you CAN do without help“ (Appendices O and P);
 - c) to student respondents in the Control group: “Read the descriptors below and tick (√) ONLY ONE box showing what your target is, or what you actually CAN do with or without help“ (see Appendix T).

Additionally, the selections that Control group respondents are expected to make support setting targets and monitoring one’s own progress (see Table 8.4 below).

Table 8.4 The checklist for identifying targets, or the skills that learners can demonstrate with or without help.

This is my target	I can now do this with help	I can now do this without help
-------------------	-----------------------------	--------------------------------

- the CEFR levels were determined per group, following the principle suggested by the authors of the checklists: if student respondents select 80% and above of the descriptors per level, this can be an indicator that they have mastered that level. If, however, subject specialists select 80% of the descriptors per level (or above), indicating the desired level for speaking ability in their context.

The CEFR level descriptors modified into actual can-do statements help the respondents recognize and express their ability more accurately. Student respondents will take the survey twice, once at the beginning of the semester, and then again at the end of the semester (subject specialist informants take the survey only once, during Phase 1). At the beginning of the semester, students will try to estimate their own speaking production and interaction skills before they receive any instructions as to how to monitor their own progress and map it on the CEFR scales. When the semester commences, students in experimental groups will receive training on how to monitor their own and the progress of their peers and map it on self-evaluation CEFR checklist. This process of familiarization with rating criteria is described in 8.2.2.2 above.

Data collected in the manner described above will be collated and statistically analyzed by the means of the following tests:

- The Sign test will be used to test and validate Hypothesis 4,
- The Pearson's Chi-Square Contingency test and Cohen's correlation measure will be applied to test and validate Hypothesis 5,
- The Mann-Whitney test will be administered to compare paired groups and test/validate Hypothesis 6.

8.2.4 End-of-semester group oral presentation task

At the end of semester, students will be able to demonstrate their speaking and presentation skills by delivering the end-of-semester group oral presentation. The first group of

students will have been practicing the same format of oral presentation from the beginning of the semester, whereas the other two groups of students will become familiar with its format and requirement immediately after the task announcement in Week 6 as per the task timeline. The task will be executed following the Task specifications and the task timeline (see Group task specifications, Chapter 7.4.3.1, Table 7.25 above).

8.2.4.1 Group presentation task – assessment

The group presentation task is a part of the formative assessment process. Two types of ratings will be provided to objectively assess the performance:

- instructor/assessor ratings, and
- peer-ratings.

8.2.4.1.1 Instructor ratings

Instructors will make use of the analytic rating scale to assess students' performance on the group speaking task (see Appendix J). The analytic rating scale includes the following criteria (see Appendix J), applied on the scale of 1 - 4:

- Group dynamics and Presentation structure
- Visuals and Audience engagement
- Non-verbal communication
- Verbal communication
- Grammar and vocabulary

The total number of points assigned to the task is 10, and assessors are instructed to divide the actual points (max. 20 pts.) in the rubric by 2 in order to obtain the total points per performance (max. 10pts.). Regardless of the provision of the summative grade, the assessment process is formative, aimed at students' enhancement of English speaking skills.

8.2.4.1.2 Peer-ratings

Students' peer-ratings reflect another difference among experimental groups regarding the execution of the oral presentation task. From the beginning of the semester, students in Group 1 will learn how to use analytic scales to evaluate their own and the performance of their peers (for the actual analytic scale used to rate peer performance on a group speaking task, see

Appendix L). Group 2 will be engaged in the same activity, but students in this group will be using holistic scales to rate their own and the performance of their peers on an individual speaking task (for the actual holistic scale used to rate peer performance on an individual speaking task, see Appendix N). Students in the Control group, however, will not be required to perform self-evaluation and evaluation of others in any of their class speaking activities.

At the end of the semester, however, the task to prepare and deliver an oral presentation will be mandatory for all students. Given that the presentations will take place during joint sessions, students in the audience will be rating the performance of their peers in order to demonstrate their ability to critically assess the work of others, which is essential in many workplaces. The rating process will require that students use rating scales in order to minimize subjectivity throughout the process exactly in the same manner they were instructed throughout the semester. In the case of the experimental groups, it should be noted that students in Group 1, who were trained to apply analytical rating scales (Appendix L), will now assess their peers' performance using a holistic rating scale to grade a group speaking task (Appendix M). On the other hand, students in Group 2, who learned how to apply a holistic scale to rate the peer performance on an individual speaking task, will use the analytical rating scale to rate the performance on the group speaking task (Appendix L). Students in experimental groups will receive the instructors' explanation on how to use either rating scale prior to the designated presentation session. Students in both groups will participate in standardization session with the instructor in the role of a moderator and volunteers in the role of group presenters providing a performance on an impromptu group speaking task. The purpose of the standardization session is to ensure inter-rater reliability among peers.

Finally, students in the Control group, who were not required to grade their own or the performance of their peers during the semester, will receive detailed instructions on how to rate their peers' performance before the presentation sessions commence. The instructor will explain the meanings of the descriptors and provide students with examples of a good/moderate/poor performance. Students will participate in standardization session with the instructor in the role of a moderator and volunteers in the role of group presenters providing a performance on an impromptu group speaking task. The purpose of the standardization session is to ensure inter-rater reliability among peers.

Data collected in the execution of the end of semester group oral presentation task will help the author validate Hypotheses 1 and 2 (see Chapter 6.2.1).

8.2.5 Final oral exam

As per the course syllabus, students sit the writing exam, followed by the final oral exam. The written exam comprises three sections: grammar, vocabulary, and reading comprehension test. After they have successfully passed the written part of the exam, students proceed to taking the oral exam. The final oral exam consists of one speaking task, weighing 10% of the total grade, based on the total grade distribution as per the English language 2 Course Syllabus document (Ekonomski fakultet, 2016).

The final oral exam comprises a single speaking task which students take individually responding to the prompt presented by the examiner. The task is presented in a short oral interview format, taking 1 minute to prepare and 1 minute to respond. In this format, a test taker reads the prompt, prepares the response based on the cues, and then responds to the interviewer's questions, in a semi-structured response format. Data collected by this summative assessment method will help the researcher validate the H3 (see Chapter 6.2.1).

8.2.5.1 Final oral exam task specifications

8.2.5.1.1 The purpose

The purpose of the speaking task in the Final Oral exam is to assess the student's ability to process a written prompt and talk about a moderately specific-purpose, business-related topic providing their own ideas and supporting them in a short talk (monologue format), followed by a short dialogue with the examiner (question-answer format).

8.2.5.1.2 Construct definition

The task is based on the *English language 2 Course Syllabus* and it sets out to assess the following (Ekonomski fakultet, 2016): overall English language comprehensibility; idea development; topic-appropriate vocabulary; grammar and pronunciation; ability to answer questions.

8.2.5.1.3 Learning outcomes

The following learning outcomes reflect the speaking ability as per the *English language 2* Course Syllabus (ibid.):

Students will be able to:

- deliver a sustained monologue
- exchange, check and confirm information
- give or seek personal views and opinions in a discussion with an interlocutor
- give descriptions
- deliver short oral presentations after short preparation
- answer impromptu questions (based on the material providing students with background knowledge of the topic)
- use topic-appropriate vocabulary

8.2.5.1.4 The characteristics of the setting of the test task

Physical setting

The task takes place in a classroom that is well lit and with the AC controlled temperature (heating in the winter season). There is one instructor desk facing 20-36 student desks (overall seating capacity of 40 - 72 students) in the room. Apart from the chairs and desks, there is one long whiteboard, and overhead projector and the screen. The task takes place in the front of the room, with the instructor and a student sitting, facing each other.

Participants

There is one student taking the exam at a time, three more students writing notes and preparing for the exam, and instructor in the role of an interlocutor. Other students (35-65) are in the room, observing, waiting for their turn. There is occasional noise coming from the students waiting for their turn; the examiner reacts when the noise occurs. Students take the exam following the alphabetical order (by last name).

Time allotment

This kind of task requires short preparation and delivery. For this reason, the time allotment is divided into two phases: 1) preparation (up to 5 minutes), and 2) delivery (2 minutes). The overall time allotted for this task is 7 minutes per test taker.

8.2.5.1.5 Instructions for responding to the task

The following set of instructions is provided to test takers (presenting groups):

The examiner:

“In this task, you are asked to talk about a business topic based on the short reading (choose one paper from the box in front of you). Read the text carefully and summarize it, and then be ready to answer three questions. Think about the text and try to guess what possible questions I may ask you about it. You have 5 minutes to prepare and we will talk about it. Please identify business-related vocabulary that we have covered in the course and use the appropriate words in your answers.”

The test takers have 5 minutes to summarize the text and anticipate the possible questions.

Example 8.5 Final oral exam speaking task

Read and summarize orally the following text (your summary should not be longer than 1 minute):

Last year over £13bn was spent on advertising in the UK and research indicates that most people will have seen 2m sales messages by the time they are 30. Advertising is big business and often acts as the interface between commerce and culture. While there are many adverts that just irritate, there are some that are very imaginative. The production costs involved in these can reach higher figures than those for the average movie. The advertisers themselves believe they are delivering an important message because they are protecting and promoting a client’s brand and extending greater choice to the consumer. [...]

Excerpt from the actual exam question, based on BEC exams (Cambridge University Press, 2002)

Once the test taker has summarized the text, the examiner asks three questions based on the text, to check comprehension and elicit students’ speaking skills.

8.2.5.1.6 Scoring method

A single examiner (the course instructor) scores the task. The assessment criteria include the following: idea development; topic-appropriate vocabulary; grammar and pronunciation;

ability to answer questions. However, since the number of students per instructor is 250, for practicality reasons the examiner does not use a rating scale. The performance on the task is rated more or less based on the overall impression, and the quality of ratings is guaranteed by the long experience in rating oral performance in English. The performance is scored on the scale from 1 to 10 and multiplied by two (total weightage of the Final oral exam is 20% of the total grade).

8.2.5.1.7 Plan for evaluating test usefulness qualities

In the current practice, there has been no plan for evaluating the overall quality of the exams. Rating reliability has, however, been ensured by standardization sessions among the instructors teaching the course.

The data collected with regards to the Final oral exam refer to exam results, expressed as numbers on the scale of 1 – 20. Data will be collected and statistically analyzed, so that the author can test and validate H1 by the means of the following statistical instruments: the Kolmogorov-Smirnov test, The Kruskal-Wallis, and the Mann-Whitney test.

8.2.6 Student perceptions survey

The final element to the research refers to all groups taking a survey aimed at investigating their perceptions about authentic tasks and evaluation criteria. Analyzing students' perceptions about assessments, Struyven et.al (2004) reviewed 36 studies dealing with a second-order perspective of assessment, i.e. the perspective not of assessment per se, but of learners' perceptions of the assessment. Their findings include several results relevant to this thesis:

- the majority of studies concerning students' perceptions are quantitative rather than qualitative (23 out of 36);
- one of the most popular quantitative methods used in research on students' perceptions on assessment is the Likert scale (in 35 out of 36 studies);
- most studies have a sample of between 101 and 200 subjects (11 out of 36);
- the reviewed studies indicate that there is a strong correlation between students' perceptions of assessment and their approaches to learning;

- learners subjected to alternative assessment methods have positive perceptions about portfolio assessment, self- and peer assessments, and simulations;
- additionally, students feel that an assessment has a positive effect “and is fair” when it: (1) relates to authentic tasks, (2) represents reasonable demands, (3) encourages students to apply knowledge to realistic contexts, (4) emphasizes the need to develop a range of skills, and is perceived to have long-term benefits” (Sambel et al, 1997 in Struyven et al., 2004: 27).

The researcher will provide an anonymous questionnaire with a set of closed-ended questions – statements, following the format of a five-point Likert scale, with two extreme attitudes, two moderate, and one neutral point (Appendix S, parts A and B). The sample includes 150 student respondents who are required to select one statement that is true to them by choosing among the following:

1. I totally disagree
2. I mostly disagree
3. I have no opinion
4. I mostly agree, and
5. I totally agree

The following statements in the survey investigate student’s perceptions of the system of evaluation and self-evaluation:

- *The feedback I was given after my presentation helped me correct my mistakes.*
- *It is important for me to know the criteria based on which my performance is judged by the instructor.*
- *I like the idea of judging my own performance by the same criteria the instructor does.*
- *I like the idea of judging my peers’ performance by the same criteria I use to judge my own performance.*

The following statements in the survey investigate students’ perceptions of the authentic tasks used in the research:

- *The tasks we were solving this semester in English language 2 classes will help me outside classroom as well.*

- *The presentation tasks helped me build my confidence when speaking in English*
- *I like tasks allowing me to choose how to solve them, e.g. by choosing a topic or preparation material for my presentation.*
- *I like tasks resembling a project or tasks requiring a group work.*

The purpose of this survey, administered at the end of the semester, is to investigate students perceptions of authentic assessment methods utilized during the research (authentic tasks and evaluation methods). Data collected in this manner will be collated and statistically analyzed by the means of the Sign test in order to help the researcher validate H3 (see Chapter 6.2.1).

After the research has been completed, all data will be triangulated and statistically analyzed for the purpose of validating the hypotheses stated in Chapter 6.2.1.

9 Testing hypotheses

9.1 Hypothesis 1

H1: The examinees who have been thoroughly trained to apply evaluation criteria demonstrate a better overall performance in the final oral exam in comparison to the examinees who have not been thoroughly trained on applying analytic and holistic scoring criteria in assessing their own and the performance of their peers.

Students in Experimental groups (1 and 2) received training on applying analytic and holistic rating scales from the beginning of the semester until its end. Additionally, Groups 1 and 2 were exposed to authentic test tasks, within the task-based approach to assessment, by which task deliverables have relevance to the TLU contexts. Students in the Control group had a short training in which they received instructions on how to apply holistic rating scales in assessing their peers' spoken production prior to the group presentation task. Their classroom activities involve pedagogical tasks derived from the syllabus approach to construct definition. As such, these tasks possess limited situational authenticity relevant to the execution of the real life tasks. According to the Sambel et al., students feel motivated to apply deep learning skills if they are exposed to authentic assessment methods, one of which is self- and peer-assessment (1997 in Struyven et al., 2004: 27). Consequently, the author assumes that the effect of familiarity with evaluation criteria by which spoken performance is judged will affect students' performance on a speaking task on the Final oral exam. To this end, a variable *Final oral exam results* is created to test the validity of H1. Additionally, Experimental groups 1 and 2 are grouped together, under an additional variable named *Task-based approach*, whereas the Control group is defined by the lack of its exposure to the task-approach in assessment. The analysis starts with testing the normality of the variable's distribution by the means of the Kolmogorov-Smirnov test of normality (see Table 9.1 below).

Table 9.1 The Kolmogorov-Smirnov test of normality for *Final oral exam results* variable

Tests of Normality		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Final oral exam results (max. 20pts.)	No	.190	50	.000	.920	50	.002
	Yes	.128	100	.000	.931	100	.000

As can be seen in Table 9.1 above, both groups of respondents are additionally defined by the *Task-based approach* variable to specify whether the students were exposed to this approach or not. Given that the sample includes no less than 50 respondents in either group, the Kolmogorov-Smirnov test is used to process data. The table also shows that the results of the Kolmogorov-Smirnov test indicate that neither group has a normal distribution (**Sig.**<0.0005). The lack of normal distribution implies that the validity of the hypothesis has to be tested by the application of the Mann-Whitney test.

Table 9.2 The Mann-Whitney test of *Final oral exam results* variable

Test Statistics ^a	Final oral exam results (max. 20pt.)
Mann-Whitney U	1670.000
Wilcoxon W	2945.000
Z	-3.339
Asymp. Sig. (2-tailed)	.001
Report	
Median	
Task-based approach	Final oral exam results (max. 20pt.)
No	14.00
Yes	16.00
Total	15.00

The Mann-Whitney test results indicate that there is a statistically significant difference in final oral exam results between students who were familiar with evaluation criteria and those who were not (**Sig**=0.001<0.05). The performance on the Final oral exam is graded on the scale of 1 to 10, and then multiplied by 2 to give the total of 20 points (equal to the total weightage of the Final oral exam in the overall grade distribution). In other words, students' performance on the task can be graded with maximum 20 points (20% in the overall grade distribution). The Mann-Whitney test indicates that the median in Experimental groups (who have been thoroughly trained to apply evaluation criteria) is 16, whereas the median in the Control group is 14. Consequently, it leads to the conclusion that Experimental groups achieved more success in the exam.

For the purpose of a more detailed data analysis, student respondents are further divided into three sub-groups, based on the module they are enrolled in. The *Study module* variable is created to process the data.

The analysis proceeds with testing the normality of the variable's distribution by the means of the Kolmogorov-Smirnov test of normality (see Table 9.3 below).

Table 9.3 The Kolmogorov-Smirnov test of normality for *Study module* variable

Study module		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	Df	Sig.	Statistic	df	Sig.
Final oral exam results (max. 20pts.)	<i>Management</i>	.180	50	.000	.886	50	.000
	<i>Accounting and Business Finance</i>	.146	50	.010	.918	50	.002
	<i>Marketing</i>	.190	50	.000	.920	50	.002

As can be seen in Table 9.3 above, both groups of respondents (Experimental groups and the Control group) are additionally defined by the *Study module* variable according to the module the students are enrolled in. Given that the sample includes no less than 50 respondents in any of the three groups of students, the Kolmogorov-Smirnov test will be used to process data. The table also shows that the Kolmogorov-Smirnov test results indicate that not even one of the three groups defined by the *Study module* variable has a normal distribution (*Management* - **Sig.**<0.0005; *Accounting and Business Finances*-**Sig.**=0.01<0.05; *Marketing*- **Sig.**<0.0005). The lack of normal distribution implies that the Kruskal-Wallis test should be used to further process data for the purpose of determining statistical difference.

Table 9.4 The Kruskal-Wallis Test for *Final oral exam* variable grouped by *Study module* variable

Test Statistics^{a,b} The Kruskal-Wallis Test		Final oral exam results (max. 20pt.)
Chi-Square		11.151
Df		2
Asymp. Sig.		.004

The Kruskal-Wallis test results indicate that there is a statistically significant difference in *Final oral exam results* among students in different modules (**Sig**=0.004<0.05). The following step is to determine among which groups a deviation occurs.

The Mann-Whitney Test is an appropriate instrument to use when comparing paired groups:

- (1) The following table indicates the difference between the groups of students enrolled in the following two modules: *Management* and *Accounting and Business Finance*(see Table 9.5 below).

Table 9.5 The Mann-Whitney test of *Final oral exam results* variable assessing the differences between *Management* and *Accounting and Business Finance*, grouped by *Study module* variable

Test Statistics ^a	Final oral exam results (max. 20pt.)
Mann-Whitney U	1250.000
Wilcoxon W	2525.000
Z	.000
Asymp. Sig. (2-tailed)	1.000

a. Grouping Variable: Study module

The Mann-Whitney test indicates that there is NO statistically significant difference in final oral exam results between students who are enrolled in *Management* module (Group 1) and those who are enrolled in *Accounting and Business Finance* (Group 2) module(**Sig.**=1.00>0.05). The Mann-Whitney test indicates that there is no statistically significant difference between the experimental groups, both exposed to the Task-based approach.

The following step is to investigate the source of discrepancy between the Experimental groups and the Control group. Again, the Mann-Whitney test will be applied to compare groups in pairs.

- (2) The following table indicates the difference between the groups of students enrolled in the following two modules: *Management* and *Marketing* (see Table 9.6 below).

Table 9.6 The Mann-Whitney test for *Final oral exam results* variable assessing the differences between *Management* and *Marketing* modules, grouped by *Study module* variable

Test Statistics ^a	Final oral exam results (max. 20pt.)
Mann-Whitney U	828.000
Wilcoxon W	2103.000
Z	-2.947
Asymp. Sig. (2-tailed)	.003

The Mann-Whitney test indicates that there is a statistically significant difference in final oral exam results between students who are enrolled in *Accounting and Business Finance* module and those who are enrolled in *Marketing* module(**Sig.**=0.003<0.05).

Finally, we will examine the difference between students enrolled in *Accounting and Business Finance* module and *Marketing* module.

(3)The following table indicates the difference between the groups of students enrolled in the following two modules: *Accounting and Business Finance* and *Marketing* (see Table 9.7 below).

Table 9.7 The Mann-Whitney test for *Final oral exam results* variable assessing the differences between *Accounting and Business Finance* and *Marketing*, grouped by *Study module* variable

Test Statistics ^a	Final oral exam results (max. 20pt.)
Mann-Whitney U	842.000
Wilcoxon W	2117.000
Z	-2.832
Asymp. Sig. (2-tailed)	.005

a. Grouping Variable: Study module

Again, the Mann-Whitney test indicates that there is a statistically significant difference in final oral exam results between students who are enrolled in *Accounting and Business Finance* and *Marketing* modules respectively(**Sig.**=0.005<0.05).

The following step in the analysis is to provide the median report of the Final oral exam results, comparing median values among the three groups of examinees (See Table 9.8 below).

Table 9.8 The median report for *Final oral exam results* variable grouped by *Study module* variable

Report Median Study module	Final oral exam results (max. 20pt.)
<i>Management</i>	17.00
<i>Accounting and Business Finances</i>	16.00
<i>Marketing</i>	14.00
Total	15.00

As can be seen in the Table 9.8 above, the median results indicate that Final oral exam results have the lowest value in the *Marketing* module ($M=14$), or in the Control group. The other two groups of students, enrolled in *Management* module (Group 1) and *Accounting and Business Finance* module (Group 2) demonstrate statistically higher results (*Management* - $M=16$; *Accounting and Business Finance*- $M=17$). The test results indicate that there is a difference in median results even between the two Experimental groups, but this difference is not statistically significant as confirmed by the Mann Whitney test (Table 9.7).

Comments on H1: The hypothesis can be fully accepted. The examinees who have been thoroughly trained to apply evaluation criteria demonstrate a better overall performance in the final oral exam in comparison to the examinees who have not been thoroughly trained on applying analytic and holistic scoring criteria in assessing their own and the performance of their peers. What is more, the median values confirm that there is no significant difference in the performance on the Final oral exam between the students enrolled in *Management* and *Accounting and Business Finance* modules respectively, indicating that their performance on the exam is similar. On the other hand, students enrolled in the Control group (*Marketing* module) demonstrate a much weaker performance, as confirmed by the median value ($M=14$). The difference in the performance between the experimental and control group can be attributed to the lack of familiarity with what constitutes a sufficient/expected response in a speaking assessment.

9.2 Hypothesis 2

H2: Performing on a task requiring that test takers should possess background knowledge related to the field of *Marketing*, the Control group demonstrates very similar results to the more successful of the two experimental groups.

Chapter 2.1.3 offers a brief overview of the testing literature with regards to the influence that background (or topical) knowledge may have on test takers' performance. In most cases, when language proficiency in general is tested, the presence of items requiring that test takers should possess background knowledge to answer correctly is seen as a source of a score contamination. In specific purpose language testing, or more specifically in specific purpose language programs, background knowledge is a part of the ability tested in an assessment. Bachman and Palmer justify this by stating that in such circumstances learners acquire not only the language but also topical knowledge related to specific academic disciplines (1996: 125). According to the research proposal, all student participants in this research are required to be enrolled in *Marketing* course, regardless of whether this is mandatory or elective for their respective modules. Students in the Control group, majoring in *Marketing*, attend a wide variety of *Marketing*-related courses, including a specialized course in *Marketing*. For students in the Experimental groups, *Marketing* is an elective course.

The requirement pertaining to the knowledge of marketing is based on the assumption that background knowledge exerts a positive influence on performance in specific purpose speaking assessments. Douglas states that subject specialist informants attribute more importance to task achievement than to linguistic accuracy in performance on a task (2000). In line with this finding, the author of the thesis endeavors to investigate if students in the Control group, who are exposed to the field of marketing more than the subjects in the experimental groups, manage to employ their background knowledge and make-up for any possible linguistic deficiencies in their performance on the oral presentation task. The task is developed based on the TLU tasks, during Phase 1 of the research. It is assessed against the following criteria (on the scale of 1-4): Group dynamics and Presentation structure; Visuals and Audience engagement; Non-verbal communication; Verbal communication; and Grammar and vocabulary. The total number of points assigned to the task is 20, and assessors are instructed to divide the actual points in the rubric by 2 in order to obtain the total points per performance (max. 10 pts.).

Oral presentation results variable is created to test the validity of H2, since this variable indicates the background knowledge of *Marketing*. The analysis starts with testing the normality of the variable's distribution by the means of The Kolmogorov-Smirnov test of normality (see Table 9.9 below).

Table 9.9 The Kolmogorov-Smirnov test of normality for *Oral presentation results* variable grouped by *Study module* variable

Tests of Normality	Study module	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Oral presentation results (max. 10pt.)	<i>Management</i>	.199	50	.000	.866	50	.000
	<i>Accounting and Business Finance</i>	.140	50	.016	.947	50	.026
	<i>Marketing</i>	.198	50	.000	.882	50	.000

Given that the sample includes no less than 50 subjects in either group defined by the *Study module* variable, the Kolmogorov-Smirnov test is used to process data. According to the results of the Kolmogorov-Smirnov test (presented in Table 9.9 above), not one of the three groups has a normal distribution for the *Oral presentation results* variable (*Management* -**Sig.** <0.0005; *Accounting and Business Finance* -**Sig.** =0.016<0.05; *Marketing* -**Sig.** <0.0005). Consequently, the Kruskal-Wallis test will be applied to test the validity of H2 (see Table 9.10 below).

Table 9.10 The Kruskal-Wallis Test for *Oral presentation results* variable grouped by *Study module* variable

Test Statistics ^{a,b}	Oral presentation results (max. 10pt.)
Chi-Square	14.040
df	2
Asymp. Sig.	.001

a. Kruskal Wallis Test

b. Grouping Variable: Study module

The results of the Kruskal-Wallis test indicate that there is a statistically significant difference with reference to the *Oral presentation results* variable applied to the three different modules (**Sig**=0.001<0.05). The following step is to determine the source of deviation among the groups, by grouping them based on the variable that indicates their *Study module*, and to assess them in pairs.

The Mann-Whitney Test is an appropriate instrument to use when comparing paired groups:

- (1) The following table indicates the difference between the groups of students enrolled in the following two modules: *Management* and *Accounting and Business Finance*(see Table 9.11 below).

Table 9.11 The Mann-Whitney test for *Oral presentation results* variable assessing the differences between *Management* and *Accounting and Business Finance* modules grouped by *Study module* variable

Test Statistics ^a	Oral presentation results (max. 10pt.)
Mann-Whitney U	864.000
Wilcoxon W	2139.000
Z	-2.715
Asymp. Sig. (2-tailed)	.007

a. Grouping Variable: Study module

The Mann-Whitney test results indicate that there is a statistically significant difference with regards to the performance on an Oral presentation task between students who are enrolled in *Management* and students enrolled in *Accounting and Business Finance* (**Sig.**=0.007<0.05).

The next step is to examine the difference between students enrolled in *Management* module and students enrolled in *Marketing*:

- (2) The following table indicates the difference between the groups of students enrolled in the following two modules: *Management* and *Marketing*

Table 9.12 The Mann-Whitney test for *Oral presentation results* variable assessing the differences between *Management* and *Marketing* modules grouped by *Study module* variable

Test Statistics ^a	Oral presentation results (max. 10pt.)
Mann-Whitney U	1102.000
Wilcoxon W	2377.000
Z	-1.055
Asymp. Sig. (2-tailed)	.291

a. Grouping Variable: Study module

The Mann-Whitney test indicates that there is NO statistically significant difference in *Oral presentation results* between the groups of students enrolled in the following two modules: *Management* and *Marketing* (**Sig.**=0.291>0.05).

The following step is to examine the difference between students enrolled in *Accounting and Business Finance* module and students enrolled in *Marketing*.

Table 9.12 The Mann-Whitney test for *Oral presentation results* variable assessing the differences between *Accounting and Business Finance* and *Marketing* modules grouped by *Study module* variable

Test Statistics ^a	Oral presentation results (max. 10pt.)
Mann-Whitney U	748.000
Wilcoxon W	2023.000
Z	-3.528
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Study module

Again, the Mann-Whitney test indicates that there is a statistically significant difference in *Oral presentation results* between the groups of students enrolled in the following two modules: *Accounting and Business Finance* and *Marketing* (**Sig.**=0.005<0,05).

The following step in the analysis is to provide the median report of the Oral presentation results, comparing median values among the three groups of examinees (See Table 9.13 below).

Table 9.13 The median report for *Final oral exam results* variable grouped by *Study module* variable

Report Median Study module	Oral presentation results (max. 10pt.)
<i>Management</i>	8.00
<i>Accounting and Business Finance</i>	7.00
<i>Marketing</i>	8.00
Total	8.00

As can be seen in the Table 9.13 above, the median results confirm the findings of the Mann-Whitney paired tests. Students enrolled in *Management* module (Group 1) and students enrolled in *Marketing* module (Control group) share the same median value (M=8). Students in the Experimental Group 2 demonstrate weaker results (M=7), compared to the other two groups.

Comments on H2: The hypothesis can be fully accepted. The performance demonstrated between the stronger of the two Experimental groups (Group 1 – *Management* module) matches the performance of the Control group (*Marketing* module). In other words, students majoring in Marketing possess specific purpose background knowledge when confronted with a test task requiring that they activate that knowledge and strategic competence. Students enrolled in the stronger of the two groups underwent a thorough training on the use of self- and peer-assessment tools, and practiced the format of the assessment throughout the assessment. Additionally, students in Group 1 learned how to utilize analytic scoring rubric making them cognizant of all the aspects of the evaluation criteria used to rate the performance on the group speaking task. Students in the weaker of the two Experimental groups attended formative assessment classes, practicing individual speaking format task. However important for the development of the ability to sustain an extended monologue, this task fails to capture the elements of performance inherent to a group effort: group dynamics, transitions, collaboration.

The results of the statistical analyses conducted to validate H2 point out the following:

- background knowledge exerts a positive influence on task performance,
- the familiarity with the task format has a positive effect on the assessment results,
- the awareness of the assessment criteria has a positive effect on task performance.

9.3 Hypothesis 3

H3: End of semester survey results indicate that more than two thirds of the examinees demonstrate positive perceptions of authentic tasks, as well as of the system of evaluation and self-evaluation that they have been exposed to.

Sambel et al. (1997) suggest that students' perceptions of assessment exert influence on their learning; if assessments are perceived as "fair" and meaningful, students employ their study techniques conducive of deep learning (in Struyven et al., 2004). Bearing in mind the findings of Sambel et al, and Struyven et al., presented in 8.2.6 above, the author created two variables to test the validity of H3:

- (1) *The positive perceptions of authentic tasks*
- (2) *Positive perceptions of the evaluation and self-evaluation system*

Variable (1) represents the mean of students' responses to the statements representing their perceptions of the authentic tasks that they have been exposed to. Those are the following statements in the survey (Appendix S):

- *The tasks we were solving this semester in English language 2 classes will help me outside classroom as well.*
- *The presentation tasks helped me build my confidence when speaking in English*
- *I like the tasks allowing me to choose how to solve them, e.g. by choosing a topic or preparation material for my presentation.*
- *I like the tasks resembling a project or tasks requiring group work.*

The responses to the statement prompts are interpreted based on the Likert scale with the following meanings: 1- Strongly Disagree, 2 – Disagree, 3 – No opinion, 4 – Agree, 5 – Strongly Agree.

Variable (2) represents the mean of students' responses to the statements representing their perceptions of the system of evaluation and self-evaluation that they have been exposed to. Those are the following statements in the survey (Appendix S):

- *The feedback I was given after my presentation helped me correct my mistakes.*
- *It is important for me to know the criteria based on which my performance is judged by the instructor.*

- *I like the idea of judging my own performance by the same criteria the instructor does.*
- *I like the idea of judging my peers' performance by the same criteria I use to judge my own performance.*

The responses to the statement prompts are interpreted based on the Likert scale with the following meanings: 1- Strongly Disagree, 2 – Disagree, 3 – No Opinion, 4 – Agree, 5 – Strongly Agree.

The analysis starts with testing the normality of the variables' distribution by the means of The Kolmogorov-Smirnov test of normality.

Table 9.14 The Kolmogorov-Smirnov test of normality for *Positive perceptions of authentic tasks* and *Positive perceptions of the evaluation and self-evaluation system* variables

Tests of Normality	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Positive perceptions of authentic tasks	.146	150	.000	.927	150	.000
Positive perceptions of the evaluation and self-evaluation system	.109	150	.000	.962	150	.000

Given that the sample includes no less than 50 respondents in either group defined by the *Positive perceptions of authentic tasks* and *Positive perceptions of the evaluation and self-evaluation system* variables, the Kolmogorov-Smirnov test is used to process data. The table shows that results of the Kolmogorov-Smirnov test indicate that neither group has a normal distribution for these two variables (**Sig**<0.0005), so the author will introduce a non-parameter technique called the Sign test in order to test the validity of H3. To this end, the author will form a *Control variable* whose value equals 4 (for all the subjects in the population).

The analysis proceeds by performing the Sign test on the following variable: *Positive perceptions of the evaluation and self-evaluation system*

The Sign test results indicate that there is NO statistically significant difference in median values between the *Positive perceptions of the evaluation and self-evaluation system* and the *Control variable* (**Sig.**=0.929>0.05). In other words, a conclusion can be drawn that more than 50% of the sample population obtained the average score of 4 or more when responding to questions 4, 5, 6 and 20 (see Table 9.15 below).

Table 9.15 The Sign test to assess the Control variable named *Positive perceptions of the evaluation and self-evaluation system*

Test Statistics	Control variable - Positive perceptions of the evaluation and self-evaluation system
The Sign test	
Z	-.089
Asymp. Sig. (2-tailed)	.929

This can be interpreted as their having positive perceptions of the system of evaluation and self-evaluation to which they were exposed throughout the semester. Similar conclusions can be drawn based on the percentiles table below:

Table 9.16 The percentiles table for *Positive perceptions of the evaluation and self-evaluation system*

Percentiles		Percentiles					
		5	10	25	50	75	90
Weighted Average (Definition 1)	Positive perceptions of the evaluation and self-evaluation system	2.7500	3.0000	3.5000	4.0000	4.2500	4.7500
Tukey's Hinges	Positive attitude towards the evaluation and self-evaluation system			3.5000	4.0000	4.2500	

It is interesting to note that no more than 5% of the population demonstrates negative attitudes to the system of evaluation and self-evaluation. In other words, no more than 5% of the student population selected answers with the average value of 2.75 or less while responding to questions 4, 5, 6 and 20.

The next step in the analysis refers to performing the Sign test on the following variable: *Positive perceptions of authentic tasks.*

Table 9.17 The Sign test to examine the Control variable named *Positive perceptions of authentic tasks*

Test Statistics^a	
The Sign Test	
	Control variable - Positive perceptions of authentic tasks
Z	-6.147
Asymp. Sig. (2-tailed)	.000

The Sign test indicates that there is a statistically significant difference between median results of the *Positive perceptions of authentic tasks variable* and the *Control variable* (**Sig**<0.0005). This difference is further examined by calculating the *Positive perceptions of authentic tasks* variable median.

Table 9.18 The median report for *Positive perceptions of authentic tasks* variable

Report
Median
Positive attitudes towards authentic tasks
4.2500

Given that the median of the variable equals $4.25 > 4$, it can be concluded that more than 50% of the student population responded to questions 3, 4, 11 and 19 answers whose value exceeds 4. In other words, more than 50% of the student population in the sample demonstrate positive attitudes towards authentic tasks. The percentile table below confirms these findings (see Table 9.19).

Table 9.19 The percentiles table for the *Positive perceptions of authentic tasks* variable

Percentiles		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Positive perceptions of authentic tasks	3.2500	3.5000	4.0000	4.2500	4.7500	4.7500	5.0000
Tukey's Hinges	Positive perceptions of authentic tasks			4.0000	4.2500	4.7500		

The percentiles table above indicates that at least 75% of the population in the sample demonstrates having positive perceptions of authentic tasks, whereas less than 5% of the population in the sample selects responses indicating their negative perceptions of authentic tasks. The graphs below will be used as a final confirmation for the conclusions drawn above (Figures 9.1-9.3):

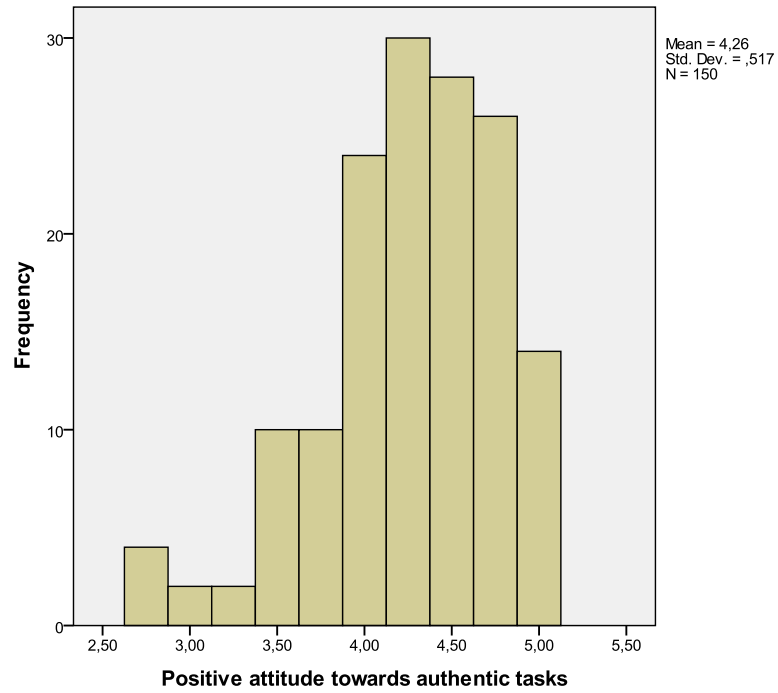


Figure 9.1 Positive perceptions of authentic tasks

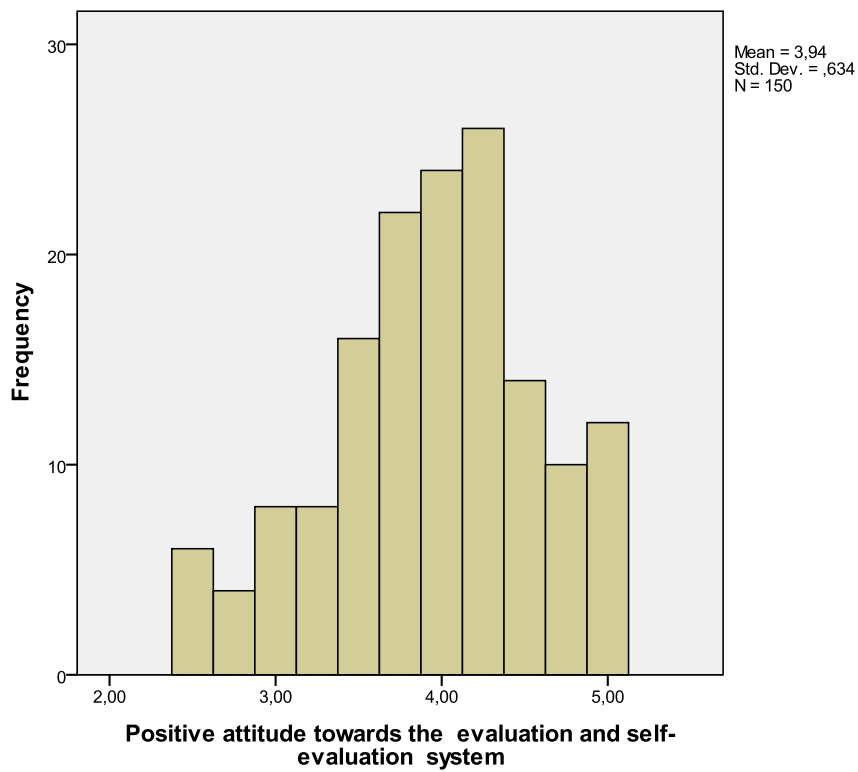


Figure 9.2 Positive perceptions of the system of evaluation and self-evaluation

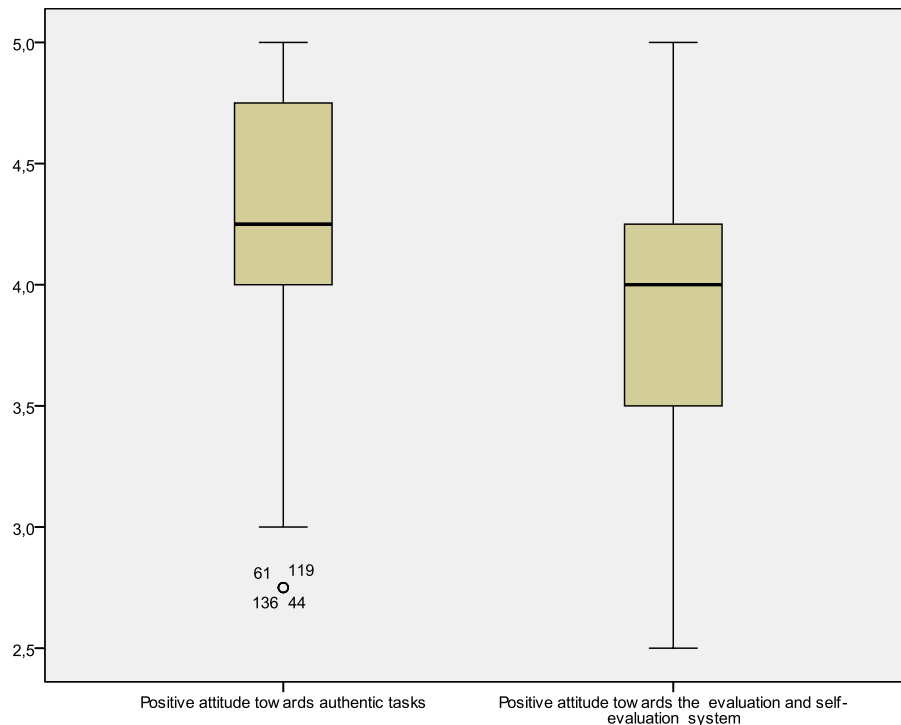


Figure 9.3 Positive perceptions

Comments on H3: The hypothesis can be fully accepted. Students in all groups demonstrate positive perceptions of authentic tasks and the system of evaluation and self-evaluation. When it comes to authentic tasks, it is interesting to observe that around three quarters of all students demonstrate having positive perceptions of the tasks which are relevant to the real life domain, i.e. the domain outside educational settings. Additionally, students express positive perceptions of the system of evaluation and self-evaluation, as more than half of the population demonstrates positive attitudes towards it. In both cases, the number of students who express negative perceptions of authentic tasks and the system of evaluation and self-evaluation does not exceed 5%, which is at the same time equal to type 1 statistical error, and as such it can be disregarded.

9.4 Hypothesis 4

H4: End of semester self-evaluation questionnaire results indicate that at least 70% of the Control group's responses provided to estimate their target skills match the responses provided at the beginning of the semester.

Subjects in the Control group were asked to respond to the "Can-do" checklist at the beginning and the end of the semester. The checklist they were provided with occasions contains shuffled descriptors mapping spoken skills on the scale ranging from A1 to B2. The C-level descriptors were intentionally excluded, since the Placement test results indicate that few students can perform at C levels at this stage of their education. Additionally, the self-evaluation checklist provided to the Control group subjects allows for planning and monitoring one's progress. The descriptors were explained to the subjects, who were instructed to make a copy of the checklists and keep it till the end of the semester, in order to map their own progress. At the end of the semester, however, the author provided them with another set of checklists, containing the same descriptors and asked them to complete the survey. The expected responses included checking only one box, indicating what students can do with someone's help, what they can do without anyone's help, or what their target is (See Table 8.4, Chapter 8.2.3.1 above). The responses were collected, ordered (by using the key for ordering descriptors, see Appendix P) and compared in order to test the validity of H4.

The self-evaluation questionnaire is presented in the form of a checklist, containing the following: 16 descriptors at A1 level (9 describing spoken interaction, and 7 describing spoken production), 23 descriptors at A2 level (13 describing spoken interaction, and 10 describing spoken production), 22 descriptors at B1 level (12 describing spoken interaction, and 10 describing spoken production), and 17 descriptors at B2 level (9 describing spoken interaction, and 8 describing spoken production). The author was interested in the subjects' ability to map their progress and review the targets they had set at the beginning of the semester (the author observed the performance of the whole group, not individuals in it, so all data represent the average response per group). In the same vein, the author endeavored to investigate if students were able to recognize the progress they have made. The reader should bear in mind that, following recommendations of the Council of Europe (2015), the attainment of a level is confirmed if 80% and more descriptors at that level are selected. However, in order to test H4,

the author compared the responses related to the target set by the subjects, assuming that during the semester of learning the majority of targets would have been met. However, in the case of the Control group, subjects were not exposed to continuous and thorough self-evaluation methods throughout the semester. Instead, they received the instructions on how to utilize the checklists at the beginning of the semester, and there was no further intervention on the part of their instructor until the end of the semester when they were asked to provide responses to the checklist again.

The aim of H4 is to investigate if students are capable of recognizing the progress they have made. If so, it can be assumed that they will have fewer target skills to select at the end of the semester. The percent to which the responses provided at the end of the semester match the responses provided at its beginning will reveal if students keep setting the same targets. If they are aware of their own progress that percent will be low, otherwise, the percent indicating the extent to which the responses overlap will be high.

The following variable was created to test the validity of H4: *The percent showing the matching between selections in the Can - do self-evaluation checklist at the beginning and the end of the semester.*

The analysis starts with testing the normality of the variable's distribution by the means of The Kolmogorov-Smirnov test of normality.

Table 9.20 The Kolmogorov-Smirnov test of normality for *The percent showing the matching between selections in the Can - do self-evaluation checklist at the beginning and the end of the semester* variable

Tests of Normality	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
The percent showing the matching between selections in the Can - do self-evaluation checklist at the beginning and the end of the semester	.184	50	.000	.788	50	.000

Given that the population in the sample includes no less than 50 respondents, the Kolmogorov-Smirnov test is used to test the variable. The findings presented in Table 9.20 above, indicate that based on the Kolmogorov-Smirnov test results, there is no normal distribution of responses (**Sig.**=<0.0005). The following step is to apply a non-parameter

technique called the Sign test in order to test the validity of the hypothesis. To this end, the author will form a Control variable to test the validity of H4.

Table 9.21 The Sign test to assess the Control variable named *Control variable-The percent showing matching between the responses in Can - do self-evaluation checklist at the beginning and at the end of semester*

Test Statistics^a The Sign Test	Control variable - The percent showing matching between the responses in Can - do self-evaluation checklist at the beginning and at the end of semester
Z Asymp. Sig. (2-tailed)	-6.207 .000

The Sign test results indicate that, at average, the percent of the responses matches in the Can-do questionnaire at the beginning and the end of the semester is statistically much different from 70% (**Sig.**=<0.0005). The median values reveal the direction of the difference.

Table 9.22 The median report for *The match between responses to self-evaluation questionnaire at the beginning and at the end of the semester*

Report Median
The percent showing matching between the responses in Can - do self-evaluation checklist at the beginning and at the end of semester
78.0000

The median report, presented in Table 9.22 above, indicates that the percentage of responses matches is significantly higher than the assumed 70%, as it equals 78%. In other words, the responses recorded at the end of the semester, when students self-evaluated their targets regarding their own speaking skills, match the same descriptors selected during the same process at the beginning of the semester.

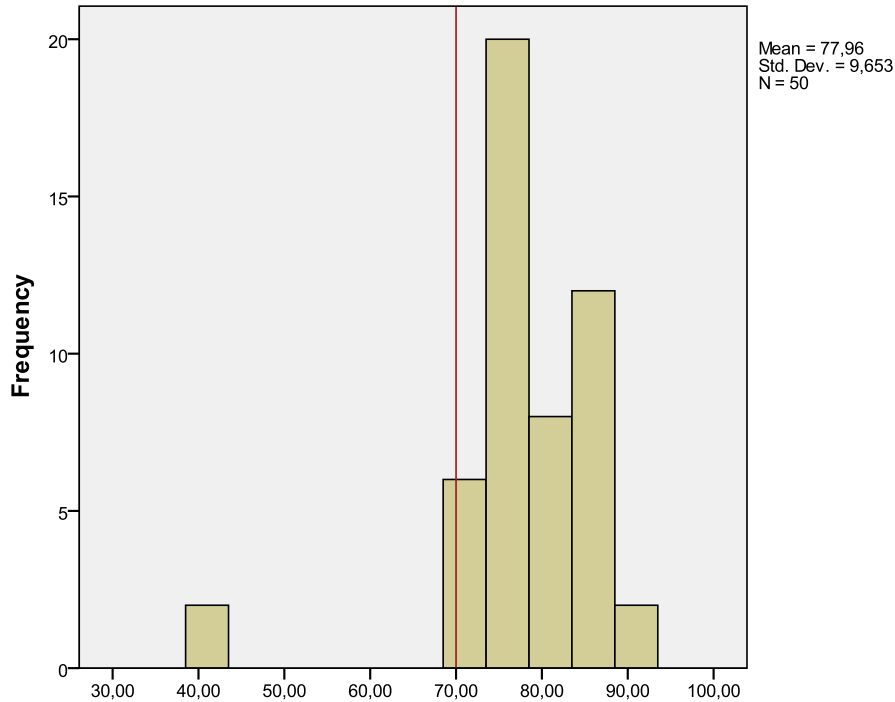


Figure 9.4 The match between the responses to self-evaluation questionnaire at the beginning and at the end of the semester

The histogram above confirms the findings (see Figure 9.4), indicating that there is a significantly higher number of matching responses, confirming that H4 can be accepted. In other words, students enrolled in *Marketing* module (Control group) selected 78% of the same target descriptors, confirming the author's assumption that their inability to recognize their own progress (corroborated by the Second placement test data, stating that the group average increased by 6.8%, i.e. from 50.13% it rose to 56.93% at the end of the semester). However, it should be noted that the author observed the whole group, not individual students in it, implying that some individual students did recognize their own progress. In addition, although the group's CEFR level did not change (it remained at B1), based on the results of the Placement test, the average point (expressed in %) increased, which can be interpreted as the sign of the group's progress.

Comments on H4: The hypothesis can be fully accepted. Students in the Control group did not receive a thorough training on self-evaluation that would help them monitor their own progress and map it correctly on the CEFR checklist. The purpose of the checklist (the one provided to the Control group, Appendix T) is to set targets related to one's own progress, but the research results imply that this process should be accompanied by clarifying the process of self-evaluation and familiarizing students with criteria by which a successful performance is judged. Additionally, it is advisable that instructors perform occasional monitoring, asking students to revisit the checklist and record the date when they have realized that they have achieved the target (this is also one of the checklist authors' recommendation).

When it comes to validating H4, the research results imply that students selected the higher number of the same target descriptors than anticipated. On the other hand, the Placement test results can be regarded as the evidence that the group has achieved progress. Consequently, students in this group are to be expected to recognize their own progress and select fewer targets that they intend to achieve in the future.

9.5 Hypothesis 5

H5: End-of-semester self-evaluation results indicate that at least half of the sample in the Experimental groups achieved progress by one CEFR level, as corroborated by the Second placement test results.

By this hypothesis, the author aims at proving that exposure to authentic test tasks and the system of evaluation and self-evaluation exerts a positive influence not only on students' progress but on their awareness of that progress as well. Students who receive a thorough training on self-evaluation become better aware of their own strengths and weaknesses helping them to set learning goals and monitor their achievement. This awareness can be detected by students' responses to the end-of-semester self-evaluation survey, indicating that they recognize their own ability to perform on higher-order speaking tasks (by selecting the corresponding descriptors on the self-evaluation grid).

The analysis starts by creating two variables random variables to help the author test H5:

- *Students achieved progress by at least one reference level as indicated by the 2nd end-of-semester self-evaluation (Variable 1)*
- *Students achieved progress by at least one reference level as indicated by the 2nd placement test results(Variable 2)*

Variable 1 is formed in order to detect students' progress at the end of the semester, as evidenced by their responses to the end-of-semester self-evaluation survey. Variable 1, on the other hand, helps the author detect students' progress at the end of the semester indicated by the 2nd placement test results. The former, being prone to subjective judgments on behalf of the respondents will be compared to the latter – objective indication of their language ability (as confirmed by the process of the Quick Placement Test Validation, according to Geranpayeh, 2003).

The random Variable 1 is in agreement with H4, and to test it, the author will use the Chi-square contingency table in order to determine the relationship between the study module and students' progress indicated by their responses to the end-of-semester self-evaluation (See Table 9.22 below).

Table 9.22 Students' progress by CEFR levels (one level up), grouped by Study module

			Students achieved progress by at least one reference level as indicated by the 2nd end-of-semester self-evaluation		Total
			No progress	Progress achieved	
Study module	<i>Management</i>	Count % within Study module	6 12.0%	44 88.0%	50 100.0%
	<i>Accounting and Business Finance</i>	Count % within Study module	20 40.0%	30 60.0%	50 100.0%
	<i>Marketing</i>	Count % within Study module	26 52.0%	24 48.0%	50 100.0%
Total	Count % within Study module	52 34.7%	98 65.3%	150 100.0%	

The Chi-square test contingency results indicate that there is a statistically significant correlation between the course module and the progress made by students for at least one reference level based on the Can-do self-evaluation checklist, from the beginning until the end of the semester (**Pearson Chi-Square Sig**<0.0005).

Table 9.23 The results of the Pearson chi-squared test

chi-Square Tests	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18.603 ^a	2	.000
Likelihood Ratio	20.379	2	.000
Linear-by-Linear Association	17.543	1	.000
N of Valid Cases	150		

a. 0 cells (0%) have expected count less than 5. The minimum expected count is 17.33.

Based on Cohen's scale, this correlation is of medium strength (**Cramer V=0,352; Sig.<0.0005**) (see Table 9.24). In other words, the percentage of the students who have made progress varies depending on the course module they are enrolled in.

Table 9.24 The strength of the correlation

Symmetric Measures		Value	Approx. Sig.
Nominal by Nominal	Phi	.352	.000
	Cramer's V	.352	.000
N of Valid Cases		150	

By observing the sample only, we can find out how the course module can be correlated to making progress during the semester. As can be seen in Table 9.22 above, based on their responses to Can-do checklist, 88% of students enrolled in the *Management* module achieved progress by one reference level. The percent of students who achieved progress in the *Accounting and Business Finance* module is smaller, but still high, and equals 60%. It is only the *Marketing* module where the progress seems to be achieved by 48%, i.e. less than half of the students enrolled in this module.

The random Variable2 is in agreement with H4, and to test it, the author will use the Chi-square contingency table in order to determine the relationship between the study module and students' progress indicated by their results on the 2nd placement test (See Table 9.25 below).

Table 9.25 Students' progress (per module) indicated by the 2nd Placement test results

			Students achieved progress by at least one reference level as indicated by the 2 nd placement test results		Total
			No progress	Progress achieved	
Study module	<i>Management</i>	Count % within Study module	0 .0%	50 100.0%	50 100.0%
	<i>Accounting and Business Finances</i>	Count % within Study module	0 .0%	50 100.0%	50 100.0%
	<i>Marketing</i>	Count % within Study module	2 4.0%	48 96.0%	50 100.0%
Total	Count % within Study module	2 1.3%	148 98.7%	150 100.0%	

To determine correlation between the course module and placement test results at the beginning and the end of the semester the author will use the Chi-square test contingency table (see Table 9.26 below).

Table 9.26 The results of the Pearson chi-squared test

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.054 ^a	2	.132
Likelihood Ratio	4.449	2	.108
Linear-by-Linear Association	3.020	1	.082
N of Valid Cases	150		

The Chi-square test contingency table indicates that there is NO statistically significant correlation between the module students are enrolled in and the progress achieved as indicated by the Placement test results at the beginning and end of the semester (**Pearson Chi-Square Sig=0.132>0.05**). In other words, the progress students achieved on the second Placement test is balanced across the modules, showing that 96-100% of the sample population made progress at the end of the semester. However, not the same conclusion can be reached based on the self-evaluation Can-do questionnaire results, where students in the Control group demonstrate that they are not aware of the progress that they have made (See Table 9.97 below). However, this finding supports H4, as concluded in Chapter 9.4 above.

Table 9.27 The result of the students' progress as indicated by the end-of-semester self-evaluation

	Students achieved progress by at least one reference level as indicated by the 2nd end-of-semester self-evaluation		Total
	No progress	Progress achieved	
No progress	2	50	52
Progress achieved	0	98	98
Total	2	148	150

This observation can be confirmed by running the Kappa agreement test, which reveals the extent to which the responses provided at the end of semester match those at its beginning (See Table 9.28 below).

Table 9.28 The Kappa test of agreement - results

Symmetric Measures		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Measure of Agreement	Kappa	.050	.034	1.955	.051
N of Valid Cases		150			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The Kappa agreement test indicates statistically significant disagreement in recording students' progress between the Can-do questionnaire and Placement test results (**Value=0.05; Sig. =0.051**). The contingency table indicates that there are as many as 50 students whose progress was captured by the Placement test results, but their responses to the Can-do questionnaire fail to confirm that.

Comments on H5: The hypothesis can be fully accepted. Students in the experimental groups, who are enrolled in Management, and Accounting and Business finance modules respectively, achieved progress as indicated by the 2nd placement test results and confirmed by responding to the end-of-semester self-evaluation survey. The comparison between the 1st and the 2nd placement test results indicates that all students in both experimental groups achieved progress. These findings were confirmed by students' self-evaluation in experimental groups, the results of which indicate that students are aware of their own progress. However, it should be noted that Group 1 and Group 2 were both observed as a single, Experimental group, whose results were contrasted against those of the Control group. To provide a complete picture of the results, and as a suggestion for further research, the author should contrast groups in pairs in order to determine how each group performs the self-evaluation task.

9.6 Hypothesis 6

H6: The highest agreement in responses to the “Can-do” survey is the one between subject specialist informants and Group 1 subjects.

Group 1 respondents, involving students enrolled in the *Management* module, were exposed to authentic test tasks and the system of evaluation and self-evaluation throughout the semester. They were trained on applying analytic assessment criteria in assessing their own and the performance of their peers. In executing speaking tasks (e.g. a group presentation), they were required to apply all the elements typical of authentic assessments (see Chapter 4.3 above) - challenge, transfer of knowledge, metacognition, accuracy, feedback, and collaboration— while working together on a joint outcome in the setting typical of a TLU situation. It is the author’s assumption that all these efforts result in students’ progress and their awareness of it. At the same time, H4 aims at investigating if students’ progress and their speaking skills stand in the agreement with the skills required by their prospective employers.

To test this hypothesis the author will introduce the fourth group of respondents named *Employer*, representing the labor market, within the *Study module* variable. The results, concerning their responses to the “Can-do” questionnaire, will be analyzed together with students’ responses to the end-of-semester self-evaluation survey. This step in the analysis is justified by essentially the same set of descriptors that all groups were required to respond to. The only difference between the checklists provided to students and those provided to the subject specialist informants lies in the respective wording of the descriptors. The former contain descriptors in the first person singular, describing students’ own skills; the latter contains descriptors in the third person plural, describing what prospective employees are expected to be able to do.

The analysis starts with testing the normality of the variable's distribution by the means of the Kolmogorov-Smirnov test of normality (see Table 9.29 below).

Table 9.29 The Kolmogorov-Smirnov test of normality

Tests of Normality		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
2 nd "Can do"	<i>Management</i>	.250	50	.000	.876	50	.000
self-evaluation	<i>Accounting and Business Finances</i>	.307	50	.000	.796	50	.000
	<i>Marketing</i>	.313	50	.000	.773	50	.000
and evaluation	<i>Employer</i>	.289	25	.000	.759	25	.000

Given that the sample includes no less than 50 respondents in each study module, the Kolmogorov-Smirnov test will be used to process data obtained from student respondents. The table shows that the Kolmogorov-Smirnov test results indicate that neither group has a normal distribution (**Sig.**=<0.0005). However in the Employer group, the sample includes the population of less than 50 respondents, so we author will rely on the Shapiro-Wilk test results. The test results indicate that there is no normal distribution in this group either. Bearing in mind the Kolmogorov-Smirnov test results, the author will proceed by applying the Mann-Whitney test to investigate the agreement in responses provided by paired groups.

The first pair of groups to compare involves Group 1 (students enrolled in the *Management* module) and *Employer*.

Table 9.30 The Mann-Whitney test results

Test Statistics ^a	2 nd "Can do" self-evaluation - CEFR level
Mann-Whitney U	587.000
Wilcoxon W	912.000
Z	-.459
Asymp. Sig. (2-tailed)	.646

The Mann-Whitney test results(see Table 9.30 above) indicate that there is NO statistically significant difference in responses to the “Can-do” questionnaire between students who are enrolled in the *Management* module and the respondents in the *Employer* group (**Sig.** =0.646>0.05). This is confirmed by the graph below (Figure 9.5), indicating that there is a considerable match between responses provided by subjects in the respective samples.

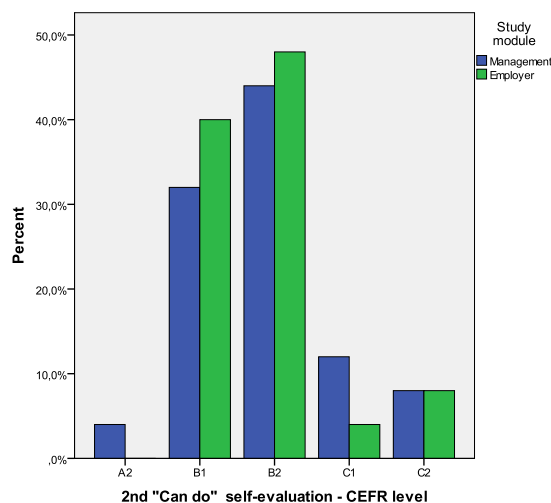


Figure 9.5 The match between responses in the following groups: *Management* module vs. *Employer* group

The following step is to examine the extent to which the responses provided by the subjects in Group 2 (students enrolled in the *Accounting and Business Finance* module) match the responses provided by the subjects in the *Employer* group.

Table 9.31 The Mann-Whitney test results

Test Statistics ^a	2 nd "Can do" self-evaluation - CEFR level
Mann-Whitney U	240.500
Wilcoxon W	1515.500
Z	-4.501
Asymp. Sig. (2-tailed)	.000

The Mann-Whitney test results(see table 9.31 above) indicate that there is a statistically significant difference in responses to the “Can-do” questionnaire between students who are enrolled in the *Accounting and Business Finance* module and the respondents in the *Employer* group (**Sig**<0.0005). This is confirmed by the graph below (see Figure 9.6), indicating that there is no match between these two groups of responses.

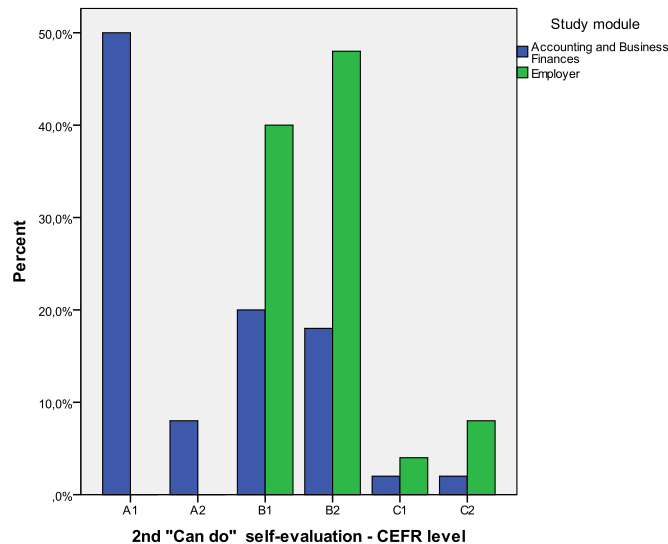


Figure 9.6 The match between responses in the following groups: *Accounting and Business Finance* module vs. *Employer* group

The final step is to examine the extent to which the responses provided by the subjects in the *Control* group (students enrolled in the *Marketing* module) match the responses provided by the subjects in the *Employer* group.

Table 9.32 The Mann-Whitney test results

Test Statistics ^a	2 nd "Can do" self-evaluation - CEFR level
Mann-Whitney U	444.000
Wilcoxon W	1719.000
Z	-2.250
Asymp. Sig. (2-tailed)	.024

The Mann-Whitney test results (see Table 9.32 above) indicate that there is a statistically significant difference in responses to the “Can-do” questionnaire between students who are enrolled in the *Marketing* module and the respondents in the *Employer* group (**Sig.** =0.024<0.0005). This is confirmed by the graph below (see Figure 9.7), indicating that there is no match between these two groups of responses.

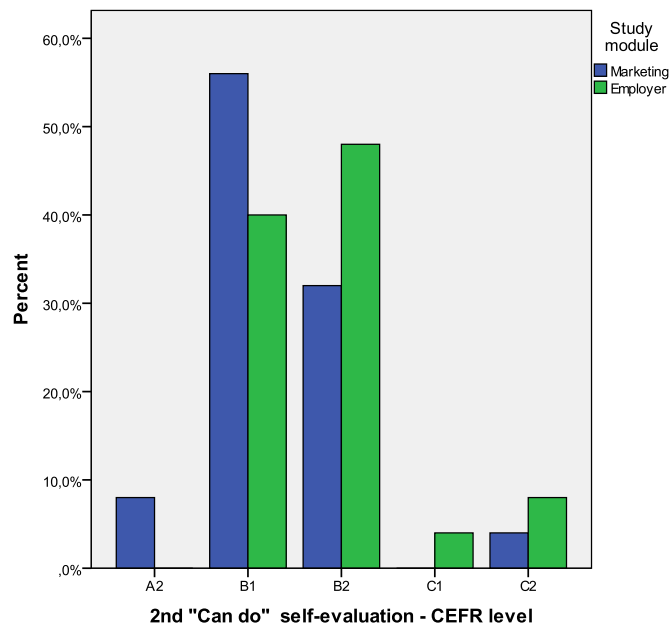


Figure 9.7 The match between responses in the following groups: Employer vs. *Marketing*

The median table below confirms the findings (see Table 9.33). The median values related to the responses provided by respective groups indicate that Group 1 and the *Employer* group share the same median value (M=3). The table indicates that other two groups of students demonstrate different median values: Group 2 (*Accounting and Business Finances*, M= 0.5), the Control group (*Marketing*, M=2). In terms of CEFR levels, the progress that Group 1 students achieved as a group (the values are taken as a group average) is evidenced in the group's awareness of the progress while responding to the end-of-semester self-evaluation survey. Consequently, the descriptors that students enrolled in the Management module selected on the occasion of the end-of-semester survey indicate that their group CEFR level shifted from B1 (at the beginning of the semester) to B2 (at the end of the semester).

To remind the reader, all groups were formed with the same group average (30 points = B1) at the beginning of the semester after they took the first placement test. At the end of the semester, all groups demonstrated progress on the second placement test. However, the most significant progress was achieved by the students in Group 1, both on the 2nd placement test and on the end-of-semester self-evaluation.

Other groups of students did achieve the progress, as evidenced in the Placement test results (see Chapter 9.5, Table 9.25).

Table 9.33 The median report

Report Median Study module	2 nd "Can do" self-evaluation and evaluation
<i>Management</i>	3.00
<i>Accounting and Business Finances</i>	.50
<i>Marketing</i>	2.00
Employer	3.00
Total	2.00

Comments on H6: The hypothesis can be fully accepted. The results of the analysis indicate that the responses provided by students enrolled in the *Management* module stand in agreement with the responses provided by the subjects in the *Employer group*. This agreement

can be interpreted by the students' awareness of their own progress, as corroborated by the 2nd placement test results. Consequently, the author interprets the agreement in these two groups of responses as the students' capability of performing well outside the educational domain, given that their speaking skills correlate with desired speaking skills in the labor market, in the real life domain. In the same vein, the reader can derive the conclusion that this is the proof of the success achieved by applying authentic test tasks and the system of evaluation and self-evaluation in the educational domain.

10 Conclusion

10.1 Introduction

This study has attempted to investigate authentic forms of assessment in testing ESP speaking skills. To achieve this objective, SP target language use speaking tasks were identified in collaboration with subject specialist informants and by the means of context-based qualitative research, helping the researcher extract speaking task characteristics in the real life domain. The TLU task characteristics were translated into test task characteristics by applying Task characteristics framework, helping the researcher develop test tasks with enhanced situational and interactional authenticity. These newly developed tasks were presented in a series of formative assessments to a group of 150 business students, enrolled in three different modules at the Faculty of Economics, along with other aspects of authentic assessments – self-evaluation, peer-evaluation, and feedback. The results obtained by assessing students' performance were collected and subjected to statistical analyses for the purpose of validating the initial hypotheses and finding answers to research questions presented in the Introduction. In Chapters 7 to 9, the research methods have been presented and discussed. In addition, Chapter 9 offers detailed findings of the statistical analyses used to validate the research hypotheses.

This chapter will conclude the thesis, and is comprised of three subsections: a summary of the main findings; an evaluation of this study in terms of its contributions to the field, its limitations and suggestions for further research.

10.2 Summary of main findings

The study presented in this study aims at investigating the influence that authentic forms of assessment have on learning and students' perceptions of their own learning in the context of ESP speaking assessment. To find answers to research questions, the author has relied on the current trends in language assessment, providing a comprehensive review of the research in assessment in Chapters 2-5. The review encompasses central issues inherent to communicative language assessment: tasks, construct definition, scores, qualities of a language test (reliability, validity, authenticity, practicality, impact/washback), authenticity (situational and interactional) and its critical elements (challenge, the focus on an outcome, transfer of knowledge,

metacognition, accuracy, environment and tools, feedback, and collaboration), target language use tasks and test tasks in the context of speaking assessment, and scoring method (including the use of analytic and holistic rating scales).

This chapter summarizes the main findings providing answers to the following research questions:

- 1) Can target language use situation tasks be used as a model for authentic classroom test tasks?
- 2) Do authentic forms of assessment exert a positive influence on students' progress?
- 3) Should background knowledge be tested in specific purpose speaking assessments?
- 4) Do authentic forms of assessment exert a positive influence on students' awareness of their own progress?
- 5) Do business students possess the language skills matching the needs of the labor market?

10.2.1 Using TLU tasks as a model for classroom test tasks

The first research question of this study was “Can target language use situation tasks be used as a model for authentic classroom test tasks?” This question has been investigated in the light of specific purpose speaking assessment, with tasks whose characteristics correspond to target language use speaking tasks.

To provide answers to the aforementioned research question, the author has applied some ideas of the so-called “grounded ethnography” technique, developed by Frenkel and Bechman in the 1980s. Originally, the approach was devised as a technique for analyzing human behavior and interactions within a narrowly defined context by videotaping the situation and analyzing it later. The purpose of the approach is to provide detailed descriptions of the situation, interaction and other important elements of the context, by having the very participants in the situation provide commentaries on the recording. Apart from the participants in TLU situations, the commentaries may include input from other relevant experts who can analyze various aspects of the situation in context. In addition to this, the commentaries from the participating parties often involve information related to indigenous assessment criteria applied by the participants in the situation. Keeping in mind a number of constraints to this approach – confidentiality issues, data

protection laws, and limited resources disabling the researcher to hire other experts, and limited applicability of the indigenous assessment criteria provided by participants in a particular context – this research relies on the commentaries provided by subject specialist informants, i.e. the experts who have a long experience of participating in the TLU situations of interest, and who know the contexts. The central role that context plays in specific purpose language testing, as acknowledged by Douglass (2000) allowed the author of the thesis to investigate the context of the TLU use and capitalize on the findings by collecting data necessary for the development of test tasks which would share the characteristics of TLU tasks. Douglas proposes hiring subject specialist informants in Selinker’s sense (1979) at an early stage of a test development project, arguing that they have the knowledge of the context and can serve as reliable judges of special purpose language performance since they know what it takes to achieve communication goals in their prospective fields. However, there are certain conditions that need to be fulfilled to this end: (1) subject specialist informants need to have a clear understanding on technical terminology used in the field, and (2) they should be prepared to respond to some language-oriented questions. Huckin and Olsen suggest that test developers and subject specialist informants should reach first reach the common ground by taking a top-down perspective of the context, and then they can proceed with analyzing other, including linguistic aspects of the context (in Douglas, 2000:99-100).

Phase 1 of the research, employing context-based research and grounded ethnography methods, with subject specialist informants feeding the research, involved data collection process that lasted from January 2015 until March 2016. The role of the subject specialist informants was to familiarize the researcher with the TLU context and help him obtain two important deliverables:

Deliverable 1 - TLU speaking task characteristics,

Deliverable 2 - the desired CEFR level for spoken production/interaction in English.

Additionally, Deliverable 1 resulted from the research employing a context-based questionnaire, consisting of two parts (Part 1, investigating a general contexts, and Part two, investigating specific speaking tasks, identified in the general part of the survey, presented in Part 1 of the questionnaire). Deliverable 2, on the other hand, was obtained through the use of “Can-do” checklists – a set of closed-ended statements, in the form of CEFR level descriptors, aimed at

investigating the desired CEFR level of prospective employees, with regards to their spoken interaction/production skills.

10.2.1.1 Deliverable 1 – TLU speaking task characteristics

The context-based research methods and subject specialist informant procedures resulted in detailed descriptions of TLU contexts in which speaking tasks in English take place. Since the research focuses on spoken performance in English, the context-based research provided details pertaining to two most common speaking tasks – a group speaking task, and an individual speaking task.

The group speaking task usually takes the form of a group presentation, typical of situations when company representatives deliver a presentation on a range of products and services, or individual products or services in front of an audience. Additionally, such tasks may involve launching a product or a service, or they can simply be dedicated to providing general details of their company. Further analyses of the responses provided by subject specialist informants revealed that these presentations normally last more than 10 minutes, and involve the collaboration among presenters, as well as the audience engagement, such as questions and activities.

The research indicates that performance on the individual speaking task requires that language users demonstrate that they are self-initiative and well prepared, that they possess background-knowledge, good interpersonal skills, as well as language knowledge required to convey the message to one or more interlocutors. In TLU contexts, individuals perform monologue-like tasks when they provide an explanation, justification, demonstration, description, or their own opinion in the course of a conversation with an interlocutor. This talk is relatively short, but still fulfills meets the requirements of extended spoken production speaking tasks, in Bachman's sense (1990).

In addition to detailed descriptions of the most common speaking tasks, the collaboration with subject specialist informants yielded a set of indigenous criteria by which language users' performance is evaluated in the real life domain. It is due to the subject-specialist input that the researcher became aware of the criteria for correctness and rating criteria that language users in TLU settings apply when they judge the speaker's performance. These sets of criteria were used

to develop rating rubrics and ensure authentic and fair evaluation of the speakers' production in the target educational setting in Phase 2.

Once the task identification process was completed, the author proceeded by translating TLU task characteristics into test task characteristics, by the means of Task characteristics framework. The following characteristics of TLU speaking tasks were analyzed: the rubric, the input, the expected response, the interaction between the input and the expected response, and the assessment. The analysis encompassed the group speaking task and the individual speaking task respectively, followed by a comparative analysis of the TLU tasks and the corresponding test tasks see Chapters 7.4.1 and 7.4.2 above). Chapter 7.4.3 summarizes the results of the analyses, providing the final result of the research conducted in Phase 1 – Test task specifications.

The resulting test task specifications document builds on the model developed by Bachman and Palmer (1996), as well as on the one developed by Douglas (2000), with an additional intervention on behalf of the author. Namely, in his attempt to bridge the gap between the educational and real life domain, the author included learning objectives into the task specifications document, so that the final task specifications documents both “captures the features of real life language use” (Green, 2014: 36) and addresses the learning outcomes outlined in the course syllabus (Ekonomski fakultet, 2016). Additionally, the model of test task specifications presented in this study enables test developers create parallel forms of the task in Fulcher's sense (2010). The following are components of the speaking test task specifications:

- the purpose of the test task,
- the definition of the construct to be measured
- the learning outcomes addressed by the construct definition
- the characteristics of the setting of the test task,
- time allotment,
- instructions for responding to the task,
- scoring method,
- plan for evaluating test usefulness qualities.

In summary to the conclusions discussed above, the author has found the answer to his first research question. Namely TLU speaking tasks can be used as a model for designing

authentic tasks for classroom use, following a thorough analysis of the context in which target language use occurs. The author has employed the principles of context-based research and grounded ethnography methods in collaboration with subject specialist informants helping him capture the characteristics of the speaking tasks performed in the real life domain. These characteristics were analyzed by the means of Task characteristics framework, and used as a model for speaking tasks in the educational domain. The Framework was used again to compare the two sets of tasks, resulting in comprehensive test task specifications document, ensuring that test tasks share situational and interactional authenticity of TLU tasks, while at the same time they meet the requirements of good testing practice.

10.2.1.2 Deliverable 2 – A desirable CEFR level for spoken interaction/production in TLU context

Deliverable 2 results from the findings based on the application of the CEFR evaluation checklists provided to subject specialist informants in Phase 1 of the research. The subjects in the survey provided their own estimations in response to a set of descriptors related to spoken interaction/production in English, identifying the CEFR level for oral performance in their respective settings. The responses were collected and analyzed, following the recommendations made by the authors of the checklists (Council of Europe, 2015). The analyses performed on the data indicate that the desired level for spoken production and interaction in English is B2. The same results have been further analyzed in Phase 2 of the research to test and validate H6, for the purpose of investigating whether any of the experimental groups possesses English language skills that match the requirements set by the representatives of the labor market.

10.2.2 Do authentic forms of assessment exert a positive influence on students' progress?

Following the research conducted during Phase 1, the author proceeded from the real life domain to the educational domain, where Phase 2 of the research evolved. The participants in Phase 2 were business students enrolled in three different study modules: *Management*, *Accounting and Business Finance*, and *Marketing*. This division into modules corresponds to the

division into two Experimental groups (Group 1 - *Management* and Group 2 - *Accounting and Business Finance*), and one Control group (*Marketing*).

All three groups took a beginning and end-of-semester placement test, as an objective measure of their proficiency in English language (see discussion in Chapter 8.2.1). Additionally, all three groups responded to “Can-do” self-evaluation survey (Chapter 8.2.3), and participated in an end-of-semester survey aimed at investigating students’ perception of authentic forms of assessment utilized during the semester (Chapter 8.2.6). All research subjects participated in end-of-semester group presentation task, developed according Test task specifications document, resulting from Phase 1 of the research (Chapter 8.2.4); and sat the final oral exam as required by all students enrolled in *English language 2* course, at the Faculty of Economics (Chapter 8.2.5).

In addition, Groups 1 and 2 were exposed to authentic forms of assessment – authentic test tasks, developed in Phase 1 of the research, and administered in formative assessment sessions throughout the semester (Chapter 8.2.2); self- and peer-assessment, accompanied by trainings on applying analytic and holistic rating scales, respectively; feedback, and collaboration. The Control group subjects, however, were not exposed to authentic tasks throughout the semester, and did not undergo continuous and thorough trainings on self-and peer-assessment; the feedback they were exposed to was limited to their performance on pedagogical, syllabus-based tasks.

To address the second research question -“Do authentic forms of assessment exert a positive influence on students’ progress?”- the author tested the following hypotheses:

H1: The examinees who have been thoroughly trained to apply evaluation criteria demonstrate a better overall performance in the final oral exam in comparison to the examinees who have not been thoroughly trained on applying analytic and holistic scoring criteria in assessing their own and the performance of their peers.

H5: End-of-semester self-evaluation results indicate that at least half of the sample in the Experimental groups achieved progress by one CEFR level, as corroborated by the Second placement test results.

The author tested the validity of H1 and H5 by the means of the following statistical instruments:

- The Kolmogorov-Smirnov Test,
- The Mann Whitney Test,

- The Kruskal-Wallis Test,
- The Pearson Chi-Square Test, and
- The Kappa Test.

According to Sambel et al. (1997), students demonstrate positive perceptions about alternative assessment methods, if they include portfolio assessment, self- and peer-assessment, and simulations (in Struyven, 2004: 27). The term *alternative assessment* can be taken as synonymous to authentic assessment, as it applies methods yielding test scores that are valuable not only in the educational context, but outside, in the real life domain. Additionally, the same study indicates that students have positive perceptions of assessment, considering it fair if it includes authentic tasks, presents students with reasonable demands, encourages them to apply knowledge to realistic contexts, and stress the importance of developing a range of skills necessary for performance outside classroom setting (ibid.). The positive perceptions of assessment methods are considered conducive of applying deep learning strategies, leading to better performance and application of knowledge across contexts (Ashford-Rowe & Herrington, 2014:18).

Discussing metacognition, as one of the critical elements of authentic assessments, Ashford-Rowe & Herrington emphasize the importance of critical reflection of one's own performance in the form of self-evaluation. Custer's view of self-evaluation is that it enhances learning (2000), since students make judgment of their own strengths and weaknesses, making strategies for improvement (Klenowski, 1995 in Ross, 2006).

In addition to self-evaluation, Topping discusses peer-evaluation as a method for making judgments about the performance of individuals of the same status, based on the set of predetermined criteria (2007 in Kerney, 2013). This, in turn, Luoma sees as an opportunity for learning as it makes students focused on the learning activity, while they are, at the same time, aware of their own learning goals.

Based on the theoretical foundations discussed here as well as in Chapters 4, 5.3.3, 8.2.3 and 8.2.6, the author developed speaking test tasks which promote situational and interactional authenticity by simulating some characteristics of the TLU tasks. The demands that students were presented with in responding to authentic tasks were reasonable and level-appropriate. At the same time, the tasks themselves were challenging, as this is one of the critical elements of

authentic assessments (see Chapter 4.3). The training on applying assessment criteria were aimed at raising students' awareness of their own performance as well as the performance of their peers, but also, at sharing the accountability for the learning process that takes place in their language classroom.

Students in Experimental groups (1 and 2) attended training on applying analytic and holistic rating scales in a series of training sessions throughout the semester. In addition to it, these students were exposed to authentic test tasks, within the task-based approach to assessment, by which task deliverables have relevance to the TLU contexts. Students in Control group had a short training in which they received instructions on how to apply holistic rating scales in assessing their peers' spoken production prior to the group presentation task. Their classroom activities involve pedagogical tasks derived from the syllabus approach to construct definition. As such, these tasks possess limited situational authenticity relevant to the execution of the real life tasks. Consequently, the author assumed that the effect of familiarity with evaluation criteria by which spoken performance is judged would result in a better performance on the Final oral exam.

By testing H1, the author obtained results which are in agreement with the initial hypothesis, finding that students in both Experimental groups achieved better results than students in Control group (see Chapter 9.1 above). Groups 1 and 2 were grouped together under the Task-based approach variable, given that both groups were exposed to task-based rather than syllabus-based approach to testing throughout the semester. By performing statistical analyses, the author reached the conclusion that experimental groups performed better than the Control group, based on the median report table (see Table 10.1 below):

Table 10.1 The median report for the Final oral exam results variable, grouped by Task-based approach

Final oral exam results (max 20 pts.)	Experimental groups	Control group
Task-based approach	Yes	No
Median	M= 16	M= 14

In addition to assessing the statistical difference based on the approach to testing spoken performance, the author performed additional analyses testing the relationship among paired groups. The median report, presented in Table 10.2 below indicates that Group 1 was the most

successful, with M=17, followed by Group 2, with M=16. The subjects in Control group demonstrate the lowest group score, with M=14.

Table 10.2 The median report for the Final oral exam results variable, grouped by Study module

Final oral exam results (max 20 pts.)	Group 1	Group 2	Control group
Median	M= 17	M= 16	M= 14

To conclude, the validation of H1 indicates that students who are familiar with evaluation criteria, and exposed to authentic test tasks, demonstrate a better general performance on summative assessments, such as the Final oral examination.

To corroborate the results obtained in the process of validation of H1, the author performed further analyses by testing the validity of H5. The results of the analyses indicate that exposure to authentic test tasks and the system of evaluation and self-evaluation exerted a positive influence on students' progress, as demonstrated by their performance on the 2nd placement test (see Chapter 9.5 above). The author analyzed the 2nd placement test results investigating progress both at an *individual* and a *group level*. The results of the analysis indicate that all individuals in the Experimental groups (*Management* and *Accounting and Business Finance*) achieved progress on the 2nd placement test, with the positive effect of shifting their CEFR level by one level up. This result is much higher than the anticipated half of the population, as hypothesized by H5. However, when it comes to the progress made at a group level, it is only Group 1 that shifted the group average from B1 to B2 level.

Table 10.2 The CEFR level results, comparison between the 1st and the 2nd placement test

Final oral exam results (max 20 pts.)	Group 1	Group 2
The 1 st placement test	B1	B1
The 2 nd placement test	B2	B1

The conclusion to be drawn is that the progress has been achieved by individuals in all groups, but it is only the results pertaining to Group 1 that achieved progress at a group level as well.

There are several reasons for the discrepancy in results between the experimental Groups 1 and 2. First, the subjects in Group 1 were exposed to group speaking tasks, requiring

collaboration and the awareness of group dynamics, accompanied by trainings on self- and peer-evaluation. Second, Group 1 subjects' awareness of the analytic rating criteria exerted a positive influence on students' progress, as documented by their participation in the end-of-semester self-evaluation survey, the results of which indicate that students in Group 1 are well aware of their own progress (88% of the sample population demonstrate the awareness of their own progress). Finally, the individual's entry levels were relatively low in the case of Group 2, with as many as 18 out of 50 students at A2 level, according to the results of the 1st placement test. Consequently, the individuals did make progress by one CEFR level, but that was not sufficient to perform as well as the individuals in Group 1, where only 12 out of 50 students scored below B1, on the 1st placement test.

In conclusion, students' exposure to authentic test tasks and authentic methods of evaluation and self-evaluation exert a positive influence on students' learning and progress. Furthermore, students demonstrate awareness of their own progress by being capable of monitoring it and expressing it by the means of CEFR level descriptors.

10.2.3 Should background knowledge be tested in specific purpose speaking assessments?

In Chapter 2.3.1, the author has discussed the place of background knowledge in language testing, stating the conflicting attitudes that language testers and validators have towards it. The main reason for the uproar against the involvement of the background or topical knowledge in construct definition is that it may contaminate the score, giving a false picture of the candidate's language ability. However, this may be the case when it comes to general ability assessment. Special purpose language ability, on the other hand, involves learning both a foreign language, in general sense, and the specific purpose terminology associated to a certain discipline, field, or profession (Bachman and Palmer, 1996). Therefore, assessing specific purpose language knowledge will depend on the purpose of the assessment and on the definition of the construct being assessed. According to Douglas (2000) and Weigle (2002), there are three possibilities with regards to the inclusion of background knowledge in the construct definition: (a) background knowledge is not included in the construct definition as it may give advantage to certain test takers over others; (b) background knowledge is included in the construct definition

when test takers are expected to have more or less similar background knowledge, such is the case in some language programs; and (c) background knowledge and language ability are defined as separate constructs and rated separately (in cases when test developers do know how homogenous the group of test takers is).

In the case of the authentic speaking task developed in Phase 1 of the research, the analytic rating scale envisages that speakers use the knowledge of marketing and specific purpose vocabulary in English. The difference among two Experimental groups and the Control group is that the latter had been exposed to marketing-related vocabulary more extensively, given that subjects in the Control group are students enrolled in *Marketing* module. Furthermore, the group presentation task requires that students possess the knowledge of marketing, in terms of norms and strategies applied when promoting a product/service. Therefore, all students participating in the research take the course in *Marketing*, either as a mandatory course (Control group), or as an elective (Group 1 and Group 2).

It is the author's assumption that background knowledge will exert a positive influence on the performance on a specific purpose language task, such as the task requiring test takers to deliver a business presentation. This assumption is grounded in previous research stating that background knowledge enables test takers to successfully complete tasks in TLU situations, where task achievement is valued as more important than language accuracy (Douglas, 2000). In the same vein, the author aims to investigate whether the possession of the background knowledge can help students enrolled in the *Marketing* module overcome weaker language knowledge when attending to the task, enabling them to deliver a satisfactory performance. To this end, the author has tested the following hypothesis:

H2: Performing on a task requiring that test takers should possess background knowledge related to the field of *Marketing*, Control group demonstrates very similar results to the more successful of the two experimental groups.

To find the answer to the third research question, the author started by identifying the stronger of the two experimental groups. The results obtained by the application of three testing instruments have indicated that students in Group 2 have demonstrated a weaker overall performance than students in Group 1 (see Table 10.3 below): the Final oral exam results, the 2nd placement test results, additionally confirmed by end-of-semester self-evaluation survey.

Table 10.3 The comparison of the results achieved by Group 1 and Group 2

Instruments	Group 1	Group 2
Final oral exam results	M = 17	M = 16
The 2 nd placement test	B2	B1
End-of-semester self-evaluation	B2	B1

The same set of instruments was used to compare the performance of Group 1 and Control group. The findings are presented in Table 10.4 below:

Table 10.4 The comparison of the results achieved by Group 1 and Control group

Instruments	Group 1	Group 2
Final oral exam results	M = 17	M = 14
The 2 nd placement test	B2	B1
End-of-semester self-evaluation	B2	B1

According to the results of the comparison, students enrolled in the Marketing module have demonstrated a weaker overall performance than students enrolled in the Management module. However, when it comes to performing on a task that requires that test takers activate and use their background knowledge, the Mann-Whitney test results indicate there is no statistically significant difference in performance between the two groups (Sig. = 0.291 > 0.05). To corroborate the results of the Mann-Whitney test, the author processed group data, searching for median values. The findings confirm the Mann-Whitney results (discussed in Chapter 9.2 above), indicating that both group's median value is 8 (see table 10.5 below).

Table 10.5 Oral presentation results (median): Group 1 and Control group

Groups	Group 1	Control group
Median	M = 8	M = 8

In summary, the analyses have indicated that students in the control group who demonstrate a weaker general performance in English, perform well on a task requiring them to

activate their specific purpose language ability, including their background knowledge of the subject matter. The author interprets this as a positive effect that background knowledge exerts on students' spoken performance, concluding that in specific purpose tests of speaking background knowledge should be included in the test construct.

10.2.4 Do authentic forms of assessment exert a positive influence on students' awareness of their own progress?

In his attempt to find the answer to this research question, the author has relied on the research into students' perceptions about assessment, presented in the works of Sambel et al. (1997) and Struyven et al. (2004), as discussed in Chapter 8.2.6 above; as well as on the research into authenticity presented in the work of Ashford-Rowe et al. (2014), discussed in Chapter 4.3. The Chapter 10.2.2 summarizes theoretical foundations pertaining to self- and peer-evaluation, as well as to students' perceptions of authentic assessment methods. Another critical element to this end is the feedback that students get during instruction and formative assessments. Brown and Abeywickrama emphasize the role of feedback in fostering future learning and revisiting both personal and course goals and objectives (2000). Luoma argues that a useful feedback should be descriptive enough to relate student performance to learning goals, so that students know which area of their performance needs improvement and what course of action can be taken. Ashford-Rowe et al. state that feedback equips students with interpersonal skills, logic and rhetoric" necessary both in pedagogic and non-pedagogic settings because it can help them determine areas of improvement, and that is "the key to progress" (Ashford-Rowe & Herrington, 2014:210).

To get an answer to the fourth research question, the author tested the following hypotheses:

H3: End of semester survey results indicate that more than two thirds of the examinees demonstrate positive perceptions of authentic tasks, as well as of the system of evaluation and self-evaluation that they have been exposed to.

H4: End of semester self-evaluation questionnaire results indicate that at least 70% of the Control group's responses provided to estimate their target skills match the responses provided at the beginning of the semester.

H5: End-of-semester self-evaluation results indicate that at least half of the sample in the Experimental groups achieved progress by one CEFR level, as corroborated by the Second placement test results.

To investigate if students' perception of authentic assessment methods are positive or not, therefore to test the validity of H3, the author conducted the survey, at the end of the semester, in which all the subjects participated. The statements presented on the Likert scale elicited responses from 1 to 5 with the following meanings: 1- Strongly Disagree, 2 – Disagree, 3 – No opinion, 4 – Agree, 5 – Strongly Agree. To test the hypothesis, the author created two variables representing the mean of students' responses (*The positive perceptions of authentic tasks*, and *Positive perceptions of the evaluation and self-evaluation system*) that underwent the Kolmogorov-Smirnov test of normality leading to the conclusion that none of the groups demonstrated a normal distribution. Consequently the analysis proceeded by the means of a non-parameter technique, called the Sign test, and the Control variable with value 4, created to test both variables.

The results of the Sign test indicate that there was no statistically significant difference in value between the subjects' responses to the variable indicating that students have positive perceptions of authentic tasks and authentic methods of evaluation and self-evaluation, and the value of the Control variable. This was confirmed by the percentiles tables (see Table 9.16, and Table 9.19, Chapter 9.3), indicating that more than 75% of the population demonstrated having positive perceptions about authentic assessment methods, whereas less than 5% of the population demonstrated negative perceptions (as confirmed in Figures 9.1 and 9.2, Chapter 9.3). The process of testing the validity of H3 proved that this hypothesis was true, and that students do have positive perceptions of authentic assessment methods employed during the semester.

Control group subjects were not exposed to continuous and thorough application of authentic assessment methods throughout the semester, but they did participate in the end-of-semester oral presentation, preceded by a short training on using peer-evaluation rating scales. In addition, they participated in the self-evaluation survey, at the beginning and in the end of the semester, responding to descriptors in checklists, prompting them to identify targets or identify what they could do with or without help. However, apart from this initial familiarization with the checklists, Control group subjects did not receive continuous trainings on self- and peer-evaluation. By testing H4, the author attempted to find the correlation between students'

awareness of their own progress and the system of evaluation and self-evaluation. If students were aware of the progress they have made, they would select fewer targets in the end-of-semester self-evaluation survey. In the same vein, if students demonstrate that they keep setting the same targets, it will indicate that they are not aware that they have made progress in the course of the semester. The author attempted to determine the progress both at an individual and a group level. When it comes to the former, the progress has been documented by the results of the 2nd placement test, stating that 96% of individual students progressed by one CEFR level. At the group level, the group average was better than on the occasion of the 1st placement test (it rose from 50.13% to 56.93%), but it remained at the same B1 level (see Chapter 9.3 above). However, by testing H4, the author endeavored to investigate whether individual students were able to demonstrate the awareness of their own progress. Having performed the test of normality and established that there was no normal distribution in the group, the author proceeded by creating the Control variable and administering the Sign test. The test results reveal that the percent by which the descriptors selected in the end-of-semester self-evaluation survey match the initial selections to a greater extent than anticipated. More specifically, the author assumed that “at least 70% of the [...] responses” would be the same, whereas the testing of H4 indicates that the actual percentage is 78%. Interpreting the results, the author concludes that students enrolled in Control group are incapable of recording their own progress.

Finally, by testing H5, as stated in Chapter 10.2.2 above, the author proved that both experimental groups achieved progress, as measured by the 2nd placement test. In addition to this, students demonstrated the awareness of their own progress, with 88% of the subjects in Group 1 and 60% of the subject in Group 2 recognizing their own progress while responding to the survey.

In conclusion, students who are trained on monitoring and assessing their own and the peer-performance demonstrate the ability to recognize their own progress.

10.2.5 Do business students possess the language skills matching the needs of the labor market?

The last research question emerged as a result of the needs analysis conducted prior to the commencement of the research. It builds on the findings of two TEMPUS projects implemented

in collaboration with labor market, aimed at bridging the gap between academia and industry (REFLESS, 2012; CONGRAD, 2014). The main findings of the needs analysis are summarized as follows:

- there are research projects and studies indicating that there is a gap between the skills and knowledge that graduate students gain in the course of their higher education and the skills and knowledge that they are expected to demonstrate in work settings;
- English language (especially oral skills) is highly valued and considered as an indicator of an overall communicative ability in this language;
- companies that are performing business operations at the territory of Serbia are mainly privately-owned;
- prospective employers often publish job advertisements online; many of the advertisements are published in English (about 30%) and require that employees be able to actively use it.

Having identified a prominent role that spoken English language has in the local labor market, the author has endeavored to investigate whether business students possess the skills that employers need. To this end, the following hypothesis was formulated:

H6: The highest agreement in responses to the “Can-do” survey is the one between subject specialist informants and Group 1 subjects.

Students enrolled in the *Management* module were exposed to authentic test tasks and the system of evaluation and self-evaluation throughout the semester. They were trained on applying analytic assessment criteria in assessing their own and the performance of their peers. In executing speaking tasks (e.g. a group presentation), they were required to apply all the elements typical of authentic assessments (challenge, transfer of knowledge, metacognition, accuracy, feedback, and collaboration – while working together on a joint outcome in the setting typical of a TLU situation (see Chapter 4.3 above). The author came to the conclusion that authentic tasks and the system of evaluation and self-evaluation exert positive influence on students’ progress and their awareness of this progress, as documented in the answers to research questions 2, 3 and 3, above.

With reference to the last research question, the author tested H6 in an attempt to investigate whether students in Group 1 possess oral English language skills that match the employers’ expectations. To this end, the fourth group of subjects was introduced, comprising of

25 subject specialist informants, representing the labor market. Given that both student respondents and labor market respondents participated in the “Can-do” survey, responding to essentially the same descriptors (see Appendices O to R), the author performed the Mann-Whitney test on grouped pairs. The Mann-Whitney test results indicate that there is no statistically significant difference between the responses provided by subjects in Group 1 and the *Employer* group (see Figure 10.1 below). The graph below indicates that, when it comes to ability to use spoken English language, subject specialist informants have relatively high expectations of their prospective employees with educational background in business.

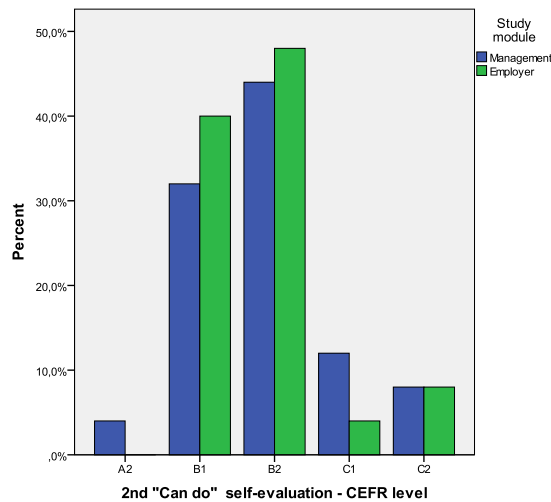


Figure 10.1 The match between responses in the following groups: *Management* module vs. *Employer* group

The green color in the bar graph demonstrates averaged responses provided in the *Employer* group, indicating that the minimum expected level is B1, with no values provided for either A1 or A2. B1 and B1 levels are the most frequent answers, coinciding with the self-evaluation results in Group 1. The median report provided for the 2nd “Can do” self-evaluation and evaluation (in the case of the *Employer* group) corroborates the Mann-Whitney test results (see Table 10.6 below).

Table 10.6 The 2nd “Can-do” evaluation and self-evaluation (median report); Employer and Group 1 (excerpt from table 9.33, Chapter 9.6)

Groups	Employer	Group 1
Median	M = 3	M = 3

Statistical analyses used to test the validity of H6 have confirmed that students enrolled in the *Management* module possess oral English language skills that match the employers' expectations.

10.3 Evaluation of the study

This chapter offers a brief outline of the main contributions and limitations identified by the author of this study.

10.3.1 Contributions

The present study makes significant contributions to theory and practice in several areas. First, this study has a theoretical significance in that it not only contributes to a better understanding of speaking assessment, but it also add knowledge to the task-based testing of ESP speaking skills in the contexts of business. In addition to this, it offers a comprehensive overview of constructs and their relevance to classroom assessment practice, in the light of specific purpose language assessment.

Second, this study contributes to ESP testing theoretically in terms of understanding of situational and interactional authenticity. By providing a set of critical elements that define authentic assessment, the study promotes development of test tasks the performance on which has value both inside and outside testing contexts.

Third, by employing the techniques pertaining to grounded ethnography approach and context-based approach, the study advocates the collaboration between test developers in educational settings and (end) test users in the real life domain in order to bridge the gap that exists between academia and industry.

Fourth, by employing the Task characteristics framework, the study has methodological significance by offering an approach that ensures systematic comparability between real life tasks and test tasks, enabling assessors to claim that the scores derived from the performance on such test tasks can generalize on the test takers' ability to perform on the corresponding real life tasks.

Fifth, the study has a pedagogical significance. The detailed investigation of students' perceptions of authentic assessment method reveals that educators should develop curricula that promote collaborative, independent and student-centered learning, while applying assessment techniques that are predominantly formative in nature. Of course, in educational settings, summative assessments are inevitable, but this study indicates that when formative methods are applied systematically and continuously, they exert a positive influence on students' performance on summative assessments.

Sixth, the study has a significant curricular contribution for the host institution – Faculty of Economic, University of Kragujevac. As it originated from the notion that English language curricula may not reflect the language needs of the end users, the study has provided guidelines for collaboration between academia and potential end users – the employers offering jobs to business graduates. The research results indicate that English language courses have a potential of educating professionals with speaking skills that match the needs of the professional domains. To this end, the study provided the following deliverables for the Faculty of Economics to consider in revisiting the English language syllabi in the next accreditation cycle – speaking test task specifications and the plan for evaluating test usefulness.

10.3.2 Limitations and suggestions for future research

While this study has made significant contributions to the fields in several academic areas as discussed above, it is not without limitations in terms of methodology and scope, which should be acknowledged. This chapter offers an overview of limitations and some suggestions for future research.

First, the method of grounded ethnography and context-based research applied in the study comes with limitations due to constraints imposed by internal company policies. The method envisages the collection of primary data by filming the participants in a communicative act. The secondary data come as a result of the primary data, in the form of comments made on the communicative act in question, enabling researchers to investigate the characteristics pertaining to their research interest. However, given that the research was not supported by companies in terms of granting a permission to film some tasks (such as presenting a business proposal, business negotiations, dealing with complaints), due to confidentiality issues, the study relies on the commentaries provided by subject specialist informants. The qualitative analysis

ensuing from this process is inevitably subjective endeavor. If other input had been applied (commentaries provided by more than one representative, assigned with different roles in a particular company), then a richer account could have been provided. The confidentiality issues do not necessarily have to affect researchers who work as in-house English instructors. Considering the fact that they already have permission to the premises, the researchers in this role could provide a valuable input into English language tasks in authentic settings.

A second limitation derives from the data, which were elicited based on a performance on two types of authentic tasks – a group presentation and a short monologue. The difficulties pertaining to practicality in terms of time, available personnel, and syllabus pacing, made it difficult for a research to employ a wider range of tasks, which would have resulted in a more comprehensive understanding of the importance of authentic assessment. The recommendation for future research pertains to diversifying tasks for data collection.

Third, certain limitations emerged as a result of validating H4. Regardless of the H4 being accepted, the author must acknowledge certain limitations of the validation process. First, the method applied fails to reveal the relationship between the responses related to the target skills and attained skills. To investigate this relationship, responses to all three columns in the checklist should be investigated to determine the nature of the relationship among them. Second, data collected in this manner can be further utilized by assessing individual progress and awareness of one's own progress, which can be taken as a suggestion for further research. Third, the author must consider other variables affecting the results of the second data collection (end-of-semester self-evaluation), such as the following: lack of motivation on behalf of students to provide honest responses, students' reliance of memory when relating their own experience to the descriptors in the checklists, the possibility that students partially agreed to the content of the descriptors, the lack of other descriptors that could better represent the students' speaking ability.

A fourth limitation of the study is in the sampling of the participants. Only three out of 6 business modules were represented, at a single Faculty of Economics in Serbia. The participants assuming the role of subject specialist informants come from 25 companies, restricted to the territory of the Municipality of Kragujevac. As a suggestion for further research, the sampling of participants in the study should expand to the territory of the whole country as it has potential relevance to Business English curricula across the country. In addition to this, a similar study should be conducted to investigate test tasks with a focus on interaction. The method of

Conversational Analysis has a lot of potential for insightful findings on the repertoire of EFL speaking in various sociolinguistic contexts.

Bibliography

- Al Azri, R. H., & Al-Rashdi, M. H. (2014). The effect of using authentic materials in teaching. *International Journal of Scientific and Technology Research*, 3(10), 249-254.
- Alderson, J. C. (1991). Bands and scores. (J. C. Alderson, & B. North, Eds.) *Language testing in the 1990s*, pp. 71-86.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.
- Ashford-Rowe, K., & Herrington, J. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education*, 32(2), 205-222.
- AUM. (2016). Dynamic Presentations. *Unpublished supplement*. American University of The Middle East.
- Austin, J. L. (1962). *How To Do Things With Words*. Oxford: OUP.
- Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453-476.
- Bachman, L. F., & Palmer, A. S. (1996). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.
- Barron's. (2006). TOEFL iBT.
- Berlak, H. (1992). The Need for a New Science of Assessment. (H. Berlak, Ed.) *Toward a New Science of Educational Testing and Assessment*.
- Boud, D. (1995). Assessment in Learning: Contradictory or Complementary? (P. Knight, Ed.) *Assessment for Learning in Higher Education*, pp. 35-48.
- British Council . (2013). *IELTS Exam Preparation Course*. Budapest: British Council English and Exams.
- Brown, G., & Craig, M. (2004,). *Assessment of Authentic Learning* . Retrieved September 15, 2015, from <http://www.coe.missouri.edu/vlib/glenn.michelle's.stuff/GLEN3MIC>

- Brown, G., & Yule, G. (1983). *Teaching the Spoken Language: an approach based on the analysis of conversational English*. Cambridge: CUP.
- Brown, H. D., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices*. White Plains, NY: Pearson Education Inc.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment* (2nd ed.). New York, NY: McGraw-Hill College.
- Brown, J. D., & Hudson, T. (1998). Alternatives in language assessment. *TESOL Quarterly*, 32(4), 653-675.
- Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.
- Bygate, M. (1987). *Speaking*. Oxford: OUP.
- Cambridge ESOL. (2002). *Quick Placement Test*. Oxford University Press.
- Cambridge University Press. (2002). *BEC Preliminary*. Cambridge: CUP.
- Cambridge University Press. (2004). *BEC Vantage 2*. Cambridge: CUP.
- Cambridge University Press. (2004). *BEC Preliminary*. Cambridge: CUP.
- Cambridge University Press. (2009). *BEC Vantage 4*. Cambridge: CUP.
- Canale, M. (1983). From communicative competence to communicative language pedagogy. (C. Richards, & R. W. Schmidt, Eds.) *Language and communication*, pp. 2-27.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor Analytic Studies*. Cambridge: Cambridge University Press.
- Castillo Losada, C. A., Insuasty, E. A., & Jaime Osorio, M. F. (2017). *The impact of authentic materials and tasks on students' communicative competence at a Colombian language school*. doi:<http://dx.doi.org/10.15446/profile.v19n1.56763>

- Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. (E. Alc ón Soler, & S. J. à, Eds.) *Intercultural Language Use and Language Learning*, pp. 41-57.
- Celce-Murcia, M., Dornyei, Z., & Thurrel, S. (1995). Communicative competence: A pedagogically motivated model with content specification. *Issues in Applied Linguistics*, 6(2), 5-35.
- Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. (L. Bachman, & A. Cohen, Eds.) *Interfaces between second language acquisition and language testing research*, 32-70.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press.
- Cohen, A. (1994). *Assessing language ability in the classroom (2nd ed.)*. Boston, MA: Heinle & Heinle.
- Cohen, A. D. (1994). *Assessing Language Ability in the Classroom*. Florence, Kentucky: Heinle & Heinle Publishers.
- CONGRAD. (2014). *Evaluacija studija i karijerni uspeh diplomiranih studenata u Srbiji i regionu*. Beograd.
- Council of Europe . (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.
- Council of Europe . (2009). *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* . Strasbourg : Council of Europe .
- Council of Europe. (2015, October 22). *Generic Checklists for Use in ELPs Designed for Language Learners Aged 15+*. Retrieved from Council of Europe: <https://www.coe.int/en/web/portfolio/the-language-biography>
- Custer, R. L. (2000). *Using Authentic Assessment in Vocational Education*. *Information Series*. No.381.ERIC Clearinghouse on Adult, Career and Vocational Education: Columbus.
- Davidson, F., & Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT: Yale University Press.

- Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
- Douglas, D., & Selinker, L. (1994). Research Methodology in context-based second language research. (E. Tarone, S. Gass, & A. Cohen, Eds.) *Methodologies for eliciting and analyzing language in context* , pp. 119-131.
- Ek, J. A. (1975). *The Treshold Level in a European Unit/Credit System for Modern Language Learning by Adults*. Strasbourg: Council of Europe.
- Ekonomski fakultet. (2016). Engleski jezik 2 . *Silabus*.
- Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, 10(3), 235-254.
- Faerch, C., & Kasper, G. (1983). *Strategies in Interlanguage Communication*. London: Longman.
- Faerch, C., & Kasper, G. (1983). *Strategies in Interlanguage Communication*. London: Longman.
- Fajgelj, S. (2009). *Psihometrija: metod i teorija psihološkog merenja* . Beograd: Centar za primenjenu psihologiju .
- Frankel, R., & Beckman, H. (1982). IMPACT: an interaction-based method for preserving and analyzing clinical transactions. (L. Pettigrew, Ed.) *Explorations in provider and patient transactions*.
- Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly*, 1(4), 253-266.
- Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.
- Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.
- Geranpayeh, A. (2003). A quick review of the English Placement Test . *Research Notes*, 8-10.
- Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing*. New York: Longman .

- Green, A. (2014). *Exploring Language Assessment and Testing: Language in Action* . New York : Routledge .
- Green, A. B. (2003). Test Impact and English for Academic Purposes: A Comparative Study in Backwash between IELTS Preparation and University Preessional Courses. *Unpublished PhD thesis*. Roehampton: University of Surrey.
- Herrington, J., & Herrington, A. (2006). Authentic conditions for Authentic Assessment: Aligning Task and Assessment. 29, 146-151. (A. Bunker, & I. Vardi, Eds.) Milperra, NSW: HERDSA.
- Huckin, T., & Olsen, L. (1984). On the use of informants in LSP discourse analysis. (A. K. Pugh, & J. M. Ulijn, Eds.) *Reading for professional purposes*, 120-129.
- Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Hughes, R. (2002). *Teaching and Researching: Speaking* (2nd Edition ed.). London: Longman.
- Hymes, D. (1972). Models of the interaction of language and social life. (J. J. Gumperz, & D. Hymes, Eds.) *Directions in sociolinguistics: The ethnography of communication*, pp. 35-71.
- Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Cambridge: Winthrop publishers.
- Jacoby, S. (1998). Science as performance: socializing scientific discourse through conference talk rehearsals. *Unpublished doctoral dissertation* . Los Angeles: University of California.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, 3(18), 213-241.
- Jacoby, S., & McNamara, T. (1999). Locating competence. *English for specific purposes*, 18(3).

- Kearney, S. (2013). Improving engagement: the use of "Authentic self- and peer- assessment for learning" to enhance the student learning experience. *Assessment and Evaluation in Higher Education*, 38(7), 875-891.
- Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education*, 2(2), 145-163.
- Laws, A. (2010). *Presentations*. Cheriton House: Heinle Cengage Learning & Summertown Publishing.
- Lebow, D. W. (1994). Authentic Activity as a Model for Appropriate Learning Activity: Implications for Emerging instructional Technologies. *Canadian Journal for Educational Communication* , 23(3), 234-444.
- Luoma, S. (2004). *Assessing Speaking* . Cambridge: Cambridge University Press.
- Lynch, B., & Davidson, F. (1994). Criterion-referenced language test development: linking curricula, teachers, and tests. *TESOL Quarterly*, 28(4), 727-743.
- McGrath, I. (2002). *Materials evaluation and design for language teaching*. Edinburgh: Edinburg University Press.
- McNamara, T. (1996). *Second Language Performance Measuring* . London and New York : Longman.
- McNamara, T., & Roever, C. (2006). Language testing: The social dimension . *Language Learning*, 1-56.
- Messick, S. (1989). Validity. (R. L. Linn, Ed.) *Educational Measurement*, pp. 13-103.
- Milanović, A., & Milanović, M. (2012). Analiza anglicizama u nazivima zanimanja u oblastima poslovne administracije i trgovine. *Srpski jezik - studije srpske i slovenske*, 17(1-2), 147-160.
- Milanović, M., & Milanović, A. (2011). Testing Listening in the Internet-based Test of English as a Foreign Language. *Nasledje*, 267-283.
- Milanović, M., & Milanović, A. (2012a). Pregled i analiza anglicizama u nazivima novijih zanimanja u poslovnim oglasima objavljenim na internetu. *Lipar*, 48, 187-201.

- Milanović, M., & Milanović, A. (2012b). Kojim jezikom govore poslodavci? O prisustvu novijih anglicizama u nazivima najtraženijih zanimanja oglašanih putem interneta. *Ekonomski horizonti*, 14(3), 177-191.
- Milanović, M., & Milanović, A. (2013). Etički problemi u testiranju jezičkih kompetencija. *Nasleđe*, X(26), 69-87.
- Milanović, M., & Milanović, A. (2014). Using the CEFR to provide test specifications for assessing vocabulary for EFL/ESL academic writing. *Nasleđe*, X(28), 57-77.
- Milanović, M., Milanović, A., & Luković, M. (2014). O odnosu konstrukata i ciljnih domena upotrebe jezika u testiranju znanja stranog jezika. *Lipar*, XV, 211-229.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centred Design*. Princetown: Educational Testing Service.
- Morrow, K. (1977). Authentic texts in ESP. (S. Holden, Ed.) *English for specific purposes*, pp. 13-15.
- Newman, F. M., & Wehlage, G. G. (1993). Five Standards of Authentic Instruction. *Educational Leadership*, 50(7), 8-12.
- Newmann, F., Marks, H., & Gamoran, A. (1996). Authentic Pedagogy and Student Performance. *American Journal of Education*, 104(4), 280-312.
- North, B. (1994). *Scales of language proficiency, a survey of some existing systems*. Strasbourg: Council of Europe CC-LANG (94) 24.
- Nunan, D. (1993). Task-based syllabus design: selecting, grading and sequencing tasks. (G. Crookes, & S. Gass, Eds.) *Tasks in a Pedagogical Context: Integrating Theory and practice*, pp. 55-68.
- Nunan, D. (2001). *Designing tasks for communicative classroom*. Cambridge: Cambridge University Press .
- Ochs, E. (1979). Transcription as theory. *Developmental Pragmatics*, 43-72.
- O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Reading, MA: Addison-Wesley Publishing Company.
- Oxford University Press. (2003). *Quick Placement Test: User Manual*. Oxford: OUP.

- Peacock, M. (1997). The effect of authentic material on the motivation of EFL learners. *ELT Journal*, 51, 2, 144-156. doi:<http://dx.doi.org/10.1093/elt/51.2.144>
- Poslovi.Infostud. (2016, March 10). *Poslovi.Infostud.* Retrieved from <http://poslovi.infostud.com>
- Powel, M. (2011). *Presenting in English: How to give successful presentations* . Cheriton House: Heinle Cengage Learning .
- Race, P., Brown, S., & Smith, B. (2005). *500 Tips on Assessment* (2nd Edition ed.). London: Routledge.
- Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.
- Rea-Dickins, P. (1991). What makes a grammar test communicative? (J. C. Alderson, & B. North, Eds.) *Language Testing in the 1990s: The Communicative Legacy*, 112-135.
- REFLESS. (2012). *Studije filologije i potrebe tržišta rada*. Beograd : Filološki fakultet .
- Ross, J. A. (2006). The Reliability, Validity, and Utility of Self-Assessment. *Practical Assessment, Research & Evaluation*, 11(10). doi:<http://pareonline.net/getvn.asp?v=11&n=10>
- Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": an exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, 23(4), 349-371.
- Selinker, L. (1979). On the use of informants in discourse analysis and language for specific purposes. *International Review of Applied Linguistics*, 17, 189-215.
- Struyven, K., Dochy, F., & Janssens, S. (2004). Students ' perceptions about evaluation and assessment in higher education. *A review 1*.
- Tonkin, A. a. (2004). Revising the IELTS speaking test. In L. E. Sheldon (Ed.), *Directions for the Future* (pp. 191-203). Bern: Peter Lang.
- Topping, K. J. (2007). Trends in peer learning. *Educational Psychology: An International Journal of Experimental Psychology*, 25(3), 635-641.
- Trigwell, K., & Prosser, M. (1991). Improving the quality of student learning: the influence of learning context and student approaches to learning on learning outcomes. *Higher Education*, 22, 251-266.

- Underhill, N. (1987). *Testing spoken language. A Handbook of oral testing techniques*. Cambridge: Cambridge University Press.
- Weir, C. (1990). *Communicative language testing*. New York: Prentice Hall.
- Weir, C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language testing*, 22(3), 1-20.
- Weir, C. J. (1993). *Understanding and developing language tests*. Englewood Cliffs, NJ: Prentice Hall.
- Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. New York: Palgrave MacMillan.
- Widdowson, H. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.
- Wiggins, G. (1993). *Assessing Student Performance - Exploring the Purpose and Limits of Testing*. San Francisco, CA: Jossey-Bass.
- Wilkins, D. A. (1976). *Notional Syllabuses*. Oxford: OUP.
- Young, R. F. (2008). *Language and interaction: An advanced resource book*. London & New York: Routledge.
- Zilvinskis, J. (2015). Using Authentic Assessment to Reinforce Student Learning in High-Impact Practices. *Assessment Update*, 27(6), pp. 7-12.

Appendices

Name:

Student ID number:

Date:

quick placement test

Version 1

The test is divided into two parts:

Part 1 (Questions 1- 40)

Part 2 (Questions 41 – 60)

Time: 30 minutes

University of Cambridge Local Examination Syndicate

Oxford University Press

Part 1

Questions 1 – 5

- Where can you see these notices?
- For questions 1 to 5, mark **one** letter **A,B** or **C** on your Answer Sheet.

1

**Please leave your
room key at Reception.**

- A in a shop
- B in a hotel
- C in a taxi

2

**Foreign money
changed here**

- A in a library
- B in a bank
- C in a police station

3

**AFTERNOON SHOW
BEGINS AT 2PM**

- A outside a theatre
- B outside a supermarket
- C outside a restaurant

4

CLOSED FOR HOLIDAYS
Lessons start again
on the 8 th January

- A at a travel agent's
- B at a music school
- C at a restaurant

5

Price per night:
£10 a tent
£5 a person

- A at a cinema
- B in a hotel
- C on a camp-site

Questions 6 – 10

- In this section you must choose the word which best fits each space in the text below.
- For questions **6** to **10**, mark **one** letter **A,B** or **C** on your Answer Sheet.

Scotland

Scotland is the north part of the island of Great Britain. The Atlantic Ocean is on the west and the North Sea on the east. Some people (6) Scotland speak a different language called Gaelic. There are (7) five million people in Scotland, and Edinburgh is (8) most famous city.

Scotland has many mountains; the highest one is called 'Ben Nevis'. In the south of Scotland, there are a lot of sheep. A long time ago, there (9) many forests, but now there are only a (10) Scotland is only a small country, but it is quite beautiful.

6 A on **B** in **C** at

7 A about **B** between **C** among

8 A his **B** your **C** its

9 A is **B** were **C** was

10 A few **B** little **C** lot

Questions 11 – 20

- In this section you must choose the word which best fits each space in the texts.
- For questions **11** to **20**, mark **one** letter **A, B, C** or **D** on your Answer Sheet.

Alice Guy Blaché

Alice Guy Blaché was the first female film director. She first became involved in cinema whilst working for the Gaumont Film Company in the late 1890s. This was a period of great change in the cinema and Alice was the first to use many new inventions, **(11)** sound and colour.

In 1907 Alice **(12)** to New York where she started her own film company. She was **(13)** successful, but, when Hollywood became the centre of the film world, the best days of the independent New York film companies were **(14)** When Alice died in 1968, hardly anybody **(15)** her name.

11A bringing **B** including **C** containing **D** supporting

12A moved **B** ran **C** entered **D** transported

13A next **B** once **C** immediately **D** recently

14A after **B** down **C** behind **D** over

15A remembered **B** realised **C** reminded **D** repeated

UFOs – do they exist?

UFO is short for 'unidentified flying object'. UFOs are popularly known as flying saucers,

(16)that is often the (17) they are reported to be. The (18)

"flying saucers" were seen in 1947 by an American pilot, but experts who studied his claim decided it had been a trick of the light.

Even people experienced at watching the sky, (19) as pilots, report seeing UFOs. In

1978 a pilot reported a collection of UFOs off the coast of New Zealand. A television

(20) went up with the pilot and filmed the UFOs. Scientists studying this

phenomenon later discovered that in this case they were simply lights on boats out fishing.

16 A because B therefore C although D so

17 A look B shape C size D type

18 A last B next C first D oldest

19 A like B that C so D such

20 A cameraman B director C actor D announcer

Questions 21 – 40

- In this section you must choose the word or phrase which best completes each sentence.
- For questions 21 to 40, mark **one** letter **A,B,C** or **D** on your Answer Sheet.

21 The teacher encouraged her studentsto an English pen-friend.

A should write **B** write **C** wrote **D** to write

22 They spent a lot of time at the pictures in the museum.

A looking **B** for looking **C** to look **D** to looking

23 Shirley enjoys science lessons, but all her experiments seem to wrong.

A turn **B** come **C** end **D** go

24 from Michael, all the group arrived on time.

A Except **B** Other **C** Besides **D** Apart

25 She her neighbour's children for the broken window.

A accused **B** complained **C** blamed **D** denied

26 As I had missed the history lesson, my friend went the homework with me.

A by **B** after **C** over **D** on

27 Whether she's a good actress or not is a of opinion.

A matter **B** subject **C** point **D** case

28 The decorated roof of the ancient palace was up by four thin columns.

A built **B** carried **C** held **D** supported

29 Would it you if we came on Thursday?

A agree **B** suit **C** like **D** fit

30 This form be handed in until the end of the week.

A doesn't need **B** doesn't have **C** needn't **D** hasn't got

31 If you make a mistake when you are writing, just it out with your pen.

- A cross B clear C do D wipe
- 32 Although our opinions on many things , we're good friends.
- A differ B oppose C disagree D divide
- 33 This product must be eaten two days of purchase.
- A by B before C within D under
- 34 The newspaper report contained important information.
- A many B another C an D a lot of
- 35 Have you considered to London?
- A move B to move C to be moving D moving
- 36 It can be a good idea for people who lead an active life to increase their of vitamins.
- A upturn B input C upkeep D intake
- 37 I thought there was a of jealousy in his reaction to my good fortune.
- A piece B part C shadow D touch
- 38 Why didn't you that you were feeling ill?
- A advise B mention C remark D tell
- 39 James was not sure exactly where his best interests
- A stood B rested C lay D centred
- 40 He's still getting the shock of losing his job.
- A across B by C over D through

Part 2

Questions 41 – 50

- In this section you must choose the word or phrase which best fits each space in the texts.
- For questions 41 to 50, mark **one** letter **A,B,C** or **D** on your Answer Sheet.

The tallest buildings - SKYSCRAPERS

Nowadays, skyscrapers can be found in most major cities of the world. A building which was many (41) high was first called a skyscraper in the United States at the end of the 19th century, and New York has perhaps the (42) skyscraper of them all, the Empire State Building. The (43) beneath the streets of New York is rock, (44) enough to take the heaviest load without sinking, and is therefore well-suited to bearing the (45) of tall buildings.

41A stages B steps C storeys D levels

42A first-rate B top-class C well-built D best-known

43A dirt B field C ground D soil

44A hard B stiff C forceful D powerful

45A weight B height C size D scale

SCRABBLE

Scrabble is the world's most popular word game. For its origins, we have to go back to the 1930s in the USA, when Alfred Butts, an architect, found himself out of (46)

..... He decided

that there was a (47) for a board game based on words and (48)

..... to

design one. Eventually he made a (49) from it, in spite of the fact that his original

(50) was only three cents a game.

46A earning B work C income D job

47A market B purchase C commerce D sale

48A took up B set out C made for D got round

49A wealth B fund C cash D fortune

50A receipt B benefit C profit D allowance

Questions 51 – 60

- In this section you must choose the word or phrase which best completes each sentence.
- For questions 51 to 60, mark **one** letter **A,B,C** or **D** on your Answer Sheet.

51 Roger's managerto make him stay late if he hadn't finished the work.

Ainsisted B warned C threatened D announced

52 By the time he has finished his week's work, John has hardly energy left for the weekend.

Aany B much C no D same

53 As the gameto a close, disappointed spectators started to leave.

- A led B neared C approached D drew
- 54 I don't rememberthe front door when I left home this morning.
- A to lock B locking C locked D to have locked
- 55 I to other people borrowing my books: they always forget to return them.
- A disagree B avoid C dislike D object
- 56 Andrew's attempts to get into the swimming team have not with much success.
- A associated B concluded C joined D met
- 57 Although Harry had obviously read the newspaper article carefully, he didn't seem to have the main point.
- A grasped B clutched C clasped D gripped
- 58 A lot of the views put forward in the documentary were open to
- A enquiry B query C question D wonder
- 59 The new college..... for the needs of students with a variety of learning backgrounds.
- A deals B supplies C furnishes D caters
- 60 I find the times of English meals very strange – I'm not used dinner at 6pm.
- A to have B to having C having D have

quick placement test

Version 2

The test is divided into two parts:

Part 1 (Questions 1- 40)

Part 2 (Questions 41 – 60)

Time: 30 minutes

University of Cambridge Local Examination Syndicate

Oxford University Press

Quick Placement Test

Part 1

Question 1 – 5

- ❖ Where can you see these notices?
- ❖ For questions 1 to 5, mark one letter **A, B** or **C** on your **Answer Sheet**.

1. YOU CAN LOOK, BUT DON'T TOUCH THE PICTURES			A	B	C
A▶ in an office	B▶ in a cinema	C▶ in a museum	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. PLEASE GIVE THE RIGHT MONEY TO THE DRIVER			A	B	C
A▶ in a bank	B▶ on a bus	C▶ in a cinema	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. NO PARKING PLEASE			A	B	C
A▶ in a street	B▶ on a book	C▶ on a table	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. CROSS BRIDGE FOR TRAINS TO EDINBURGH			A	B	C
A▶ in a bank	B▶ in a garage	C▶ in a station	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. KEEP IN A COLD PLACE			A	B	C
A▶ on clothes	B▶ on furniture	C▶ on food	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Question 6 –10

- ❖ In this section you must choose the word which best fits each space in the text below.
- ❖ For questions 6 to 10, mark **one** letter **A**, **B**, or **C** on your Answer Sheet

THE STARS

There are millions of stars in the sky. If you look **(6)**.....the sky on a clear night, it is possible to see about 3000 stars. They look small, but they are really **(7)**.....big hot balls of burning gas. Some of them are huge, but others are much smaller, like our planet Earth. The biggest stars are very bright, but they only live for a short time. Every day new stars **(8)**.....born and old stars die. All the stars are very far away. The light from the nearest star takes more **(9)**.....four years to reach Earth. Hundreds of years ago, people **(10)**.....stars, like the North Star, to know which direction to travel in. Today you can still see that star.

6.			A	B	C
A ▶ at	B ▶ up	C ▶ on	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.			A	B	C
A ▶ very	B ▶ too	C ▶ much	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.			A	B	C
A ▶ is	B ▶ be	C ▶ are	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9.			A	B	C
A ▶ that	B ▶ of	C ▶ than	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.			A	B	C
A ▶ use	B ▶ used	C ▶ using	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Question 11 - 15

- ❖ In this section you must choose the word which best fits each .space in the texts.
- ❖ For questions **11** to **20**, mark one letter **A**, **B**, **C** or **D** on your Answer Sheet.

Good smilies ahead for young teeth

Older Britons are the worst in Europe when it comes to keeping their teeth. But British youngsters **(11)**.....more to smile about because **(12)**.....teeth are among the best. Almost 80% of Britons over 65 have lost all ore some **(13)**.....their teeth according to a World Health Organisation survey. Eating too **(14)**.....sugar is part of the problem. Among **(15)**....., 12-year-olds have on average only three missing, decayed or filled teeth.

11.				A	B	C	D
A▶ getting	B▶ got	C▶ have	D▶ having	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12.				A	B	C	D
A▶ their	B▶ his	C▶ them	D▶ theirs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13.				A	B	C	D
A▶ from	B▶ of	C▶ among	D▶ between	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.				A	B	C	D
A▶ much	B▶ lot	C▶ many	D▶ deal	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15.				A	B	C	D
A▶ person	B▶ people	C▶ children	D▶ family	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Question 16 - 20

Christopher Columbus and the New World

On August 3, 1492, Christopher Columbus set sail from Spain to find a new route to India, China and Japan. At this time most people thought you would fall off the edge of the world if you sailed too far. Yet sailors such as Columbus had seen how a ship appeared to get lower and lower on the horizon as it sailed away. For Columbus this **(16)**.....that the world was round. He **(17)**.....to his men about the distance travelled each day. He did not want them to think that he did not **(18)**.....exactly where they were going. **(19)**....., on October 12, 1492, Columbus and his men landed on a small island he named San Salvador. Columbus believed he was in Asia, **(20)**.....he was actually in the Caribbean.

16.				A	B	C	D
A ► made	B ► pointed	C ► was	D ► proved	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17.				A	B	C	D
A ► lied	B ► told	C ► cheated	D ► asked	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
18.				A	B	C	D
A ► find	B ► know	C ► think	D ► expect	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19.				A	B	C	D
A ► Next	B ► Secoundly	C ► Finally	D ► Once	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
20.				A	B	C	D
A ► as	B ► but	C ► because	D ► if	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Question 21 - 30

- ❖ In this section you must choose the word or phrase which best completes each sentence.
- ❖ For questions 21 to 40, mark one letter A, B, C or D on your Answer Sheet.

21. The children won't go to sleep.....we leave a light on outside their bedroom.				A	B	C	D
A ▶ except	B ▶ otherwise	C ▶ unless	D ▶ but	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. I'll give you my spare keys in case you.....home before me.				A	B	C	D
A ▶ would get	B ▶ got	C ▶ will get	D ▶ get	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. My holiday in Paris gave me a great.....to improve my French accent.				A	B	C	D
A ▶ occasion	B ▶ chance	C ▶ hope	D ▶ possibility	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24. The singer ended the concert.....her most popular song.				A	B	C	D
A ▶ by	B ▶ with	C ▶ in	D ▶ as	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25. Because it had not rained for several months, there was a.....of water.				A	B	C	D
A ▶ shortage	B ▶ drop	C ▶ scare	D ▶ waste	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
26. I've always.....you as my best friend.				A	B	C	D
A ▶ regarded	B ▶ thought	C ▶ meant	D ▶ supposed	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
27. She came to live her.....a month ago.				A	B	C	D
A ▶ quite	B ▶ beyond	C ▶ already	D ▶ almost	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
28. Don't make such a.....! The dentist is only going to look at your teeth.				A	B	C	D
A ▶ fuss	B ▶ trouble	C ▶ worry	D ▶ reaction	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
29. He spent a long time looking for a tie which.....with his new shirt.				A	B	C	D
A ▶ fixed	B ▶ made	C ▶ went	D ▶ wore	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
30. Fortunately,.....from a bump on the head, she suffered no serious injuries from her fall.				A	B	C	D
A ▶ other	B ▶ except	C ▶ besides	D ▶ apart	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Question 31 – 40

31. She had changed so much that..... anyone recognised her.				A	B	C	D
A▶ almost	B▶ hardly	C▶ not	D▶ nearly				
32.teaching English, she also writes children's books.				A	B	C	D
A▶ Moreover	B▶ As well as	C▶ In addition	D▶ Apart				
33. It was clear that the young couple were..... of taking charge of the restaurant.				A	B	C	D
A▶ responsible	B▶ reliable	C▶ capable	D▶ able				
34. The book..... of ten chapters, each one covering a different topic.				A	B	C	D
A▶ comprises	B▶ includes	C▶ consists	D▶ contains				
35. Mary was disappointed with her new shirt as the colour..... very quickly.				A	B	C	D
A▶ bleached	B▶ died	C▶ vanished	D▶ faded				
36. National leaders from all over the world are expected o attend the.....meeting.				A	B	C	D
A▶ peak	B▶ summit	C▶ top	D▶ apex				
37. Jane remained calm when she won the lottery andabout her business as if nothing had happened.				A	B	C	D
A▶ came	B▶ brought	C▶ went	D▶ moved				
38. I suggest we..... outside the stadium tomorrow at 8.30.				A	B	C	D
A▶ meeting	B▶ meet	C▶ met	D▶ will meet				
39. My remarks were..... as a joke, but she was offended by them.				A	B	C	D
A▶ pretended	B▶ thought	C▶ meant	D▶ supposed				
40. You ought to take up swimming for the..... of your health.				A	B	C	D
A▶ concern	B▶ relief	C▶ sake	D▶ cause				

Part 2

Questions 41 – 45

- ❖ In this section you must choose the word which best fits each space in the texts.
- ❖ For questions 41 to 45, mark one letter **A**, **B**, **C** or **D** on your Answer Sheet.

CLOCKS

The clock was the first complex mechanical machinery to enter the home, (41).....it was too expensive for the (42).....person until the 19th century, when (43).....production techniques lowered the price. Watches were also developed, but they (44).....luxury items until 1868, When the first cheap pocket watch was designed in Switzerland. Watches later became (45).....available, and Switzerland became the world's leading watch manufacturing centre for the next 100 years.

41.				A	B	C	D
A ▶ despite	B ▶ although	C ▶ otherwise	D ▶ average	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
42.				A	B	C	D
A ▶ average	B ▶ medium	C ▶ general	D ▶ common	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
43.				A	B	C	D
A ▶ vast	B ▶ large	C ▶ wide	D ▶ mass	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
44.				A	B	C	D
A ▶ lasted	B ▶ endured	C ▶ kept	D ▶ remained	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
45.				A	B	C	D
A ▶ mostly	B ▶ chiefly	C ▶ greatly	D ▶ widely	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Questions 46 - 50

Dublin City Walks

What better way of getting to know a new city than by walking around it? Whether you choose the Medieval Walk, which will **(46)**.....you to the 1000 years ago, find out about the more **(47)**.....history of the city on the Eighteenth Century Walk, or meet the ghosts of Dublin's many writers on The Literary Walk, we know you will enjoy the experience.

Dublin City Walks **(48)**.....twice daily. Meet your guide at 10.30 a.m. or 2.30 p.m. at the Tourist Information Office. No advance **(49)**.....is necessary. Special **(50)**.....are available for families, children and parties of more than ten people.

46.				A	B	C	D
A ► introduce	B ► present	C ► move	D ► show	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
47.				A	B	C	D
A ► near	B ► late	C ► recent	D ► close	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
48.				A	B	C	D
A ► take place	B ► occur	C ► work	D ► function	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
49.				A	B	C	D
A ► paying	B ► reserving	C ► warning	D ► booking	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
50.				A	B	C	D
A ► funds	B ► costs	C ► fees	D ► rates	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Question 51– 60

- ❖ In this section you must choose the word or phrase which best completes each sentence.
- ❖ For questions 51 to 60, mark one letter **A**, **B**, **C** or **D** on your Answer Sheet.

51. If you're not too tired we could have a.....of tennis after lunch. **A** **B** **C** **D**

A▶ match **B**▶ play **C**▶ game **D**▶ party

52. Don't you get tired.....watching TV every nigh? **A** **B** **C** **D**

A▶ with **B**▶ by **C**▶ of **D**▶ at

53. Go on, finish the dessert. It needs.....up because it won't stay fresh until. **A** **B** **C** **D**

A▶ eat **B**▶ eating **C**▶ to eat **D**▶ eaten

54. We're not used to.....invited to very formal occasions. **A** **B** **C** **D**

A▶ be **B**▶ have **C**▶ being **D**▶ having

55. I'd rather we.....meet this evening, because I'm very tired. **A** **B** **C** **D**

A▶ wouldn't **B**▶ shouldn't **C**▶ hadn't **D**▶ didn't

56. She obviously didn't want to discuss the matter so I didn't.....the point. **A** **B** **C** **D**

A▶ maintain **B**▶ chase **C**▶ follow **D**▶ pursue

57. Anyone.....after the start of the play is not allowed in until the interval. **A** **B** **C** **D**

A▶ arrives **B**▶ has arrived **C**▶ arriving **D**▶ arrived

58. This new magazine iswith interesting stories and useful information. **A** **B** **C** **D**

A▶ full **B**▶ packed **C**▶ thick **D**▶ compiled

59. The restaurant was far too noisy to be.....to relaxed conversation. **A** **B** **C** **D**

A▶ conducive **B**▶ suitable **C**▶ practical **D**▶ fruitful

60. In this branch of medicine, it is vital toopen to new ideas. **A** **B** **C** **D**

A▶ stand **B**▶ continue **C**▶ hold **D**▶ remain

Appendix C: Key to Quick placement test Versions 1 and 2

Quick Placement Test Version 1

Key

1	B
2	B
3	A
4	B
5	C
6	B
7	A
8	C
9	B
10	A
11	B
12	A
13	C
14	D
15	A

16	A
17	B
18	C
19	D
20	A
21	D
22	A
23	D
24	D
25	C
26	C
27	A
28	C
29	B
30	C

31	A
32	A
33	C
34	D
35	D
36	D
37	D
38	B
39	C
40	C
41	C
42	D
43	C
44	A
45	A

46	B
47	A
48	B
49	D
50	C
51	C
52	A
53	D
54	B
55	D
56	D
57	A
58	C
59	D
60	B

Points	ALTE		CEFR	Relevance to Business English levels
	Level	Description		
0-10	0.1	Beginner		
11-17	0.2	Breakthrough	A 1	
18-29	1	Elementary	A2	
30-39	2	Lower-Intermediate	B1	
40-47	3	Upper – Intermediate	B 2	BEC P Business English Certificate Preliminary
48-54	4	Advanced	C1	BEC V Business English Certificate Vantage
55-60		Very Advanced	C 2	BEC H Business English Certificate Higher

Quick Placement Test Version 2

Key

1	C
2	A
3	A
4	B
5	A
6	B
7	A
8	C
9	C
10	A
11	B
12	A
13	C
14	D
15	B

16	B
17	A
18	A
19	C
20	A
21	D
22	A
23	D
24	D
25	C
26	C
27	A
28	C
29	B
30	A

31	B
32	B
33	C
34	A
35	A
36	D
37	D
38	B
39	C
40	A
41	C
42	D
43	C
44	A
45	B

46	A
47	B
48	A
49	D
50	C
51	C
52	A
53	D
54	B
55	B
56	D
57	A
58	C
59	D
60	C

Points	ALTE		CEFR	Relevance to Business English levels
	Level	Description		
0-10	0.1	Beginner		
11-17	0.2	Breakthrough	A 1	
18-29	1	Elementary	A2	
30-39	2	Lower-Intermediate	B1	
40-47	3	Upper – Intermediate	B 2	BEC P Business English Certificate Preliminary
48-54	4	Advanced	C1	BEC V Business English Certificate Vantage
55-60		Very Advanced	C 2	BEC H Business English Certificate Higher

Appendix D: Information statement and Consent form (in English)

UNIVERSITY OF KRAGUJEVAC

FACULTY OF ECONOMICS

INFORMATION STATEMENT AND CONSENT FORM

Research Title: Investigating authentic forms of assessment in testing English for specific purpose speaking skills (Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику струке)

Researcher: Milan Milanović, English language instructor, doctoral student

Dear student,

You are kindly invited to participate in the research entitled *Investigating authentic forms of assessment in testing English for specific purpose speaking skills*, the objective of which is to determine the level of authenticity of test tasks used to assess your spoken English skills at the Faculty of Economics. The research project aims at determining whether the English skills you develop at the Faculty correspond to the labor market requirements that you are expected to meet upon graduation.

Should you give your consent to participate in this research, you will be subjected to, not only the learning activities outlined in the English language 2 Course Syllabus, but more intensive and varied language assessment methods, which will have NO negative washback on your performance. Also note that all information or personal details gathered in the course of the study are confidential. If you decide to participate, you are free to withdraw from further participation in the research at any time without having to give a reason and without consequence.

I, _____ (student ID number _____), confirm by signing that I voluntarily agree to participate in this research. At the same time, I confirm that I understand my role and duties in the realization of the study. In addition, I confirm that I know that I can withdraw from further participation in the research at any time without having to give a reason and without consequence.

In Kragujevac, _____

Appendix E: Information statement and Consent form (in Serbian)

УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ

ЕКОНОМСКИ ФАКУЛТЕТ

ИЗЈАВА О НАМЕРИ ИСТРАЖИВАЧА И ИЗЈАВА О ДОБРОВОЉНОМ УЧЕШЋУ У ИСТРАЖИВАЧКОМ ПРОЈЕКТУ

Назив истраживања: Investigating authentic forms of assessment in testing English for specific purpose speaking skills (Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику струке)

Истраживач: Милан Милановић, професор енглеског језика, докторанд

Поштовани студенте,

позвани сте да учествујете у пројекту под називом *Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику струке*, који има за циљ да утврди колико су аутентичне постојеће методе тестирања усмених језичких вештина на енглеском језику који учите на Економском факултету. Пројекат, такође, има за циљ да утврди да ли су говорне вештине које стичете учећи енглески језик на факултету управо оне вештине које су вам потребне на тржишту рада након дипломирања.

Уколико пристанете да учествујете у истраживачком пројекту, осим у активностима предвиђеним силабусом предмета Енглески језик 2, бићете подвргнути интензивнијим и разноликијим методама провере језичког знања, које неће негативно утицати на ваше постигнуће. Сви лични подаци, као и подаци везани за ваш успех и постигнуће прикупљени за потребе истраживања су анонимни. Уколико одлучите да помогнете у спровођењу истраживања, а одлучите да се повучете из истог, слободни сте то да учините у било ком тренутку, без образложења и последица.

Ја, _____ (бр. индекса _____), потписујем да добровољно пристајем да учествујем у овом истраживању. Истовремено потврђујем даразумем своју улогу и обавезе у реализацији истог. Осим тога, потврђујем да ми је познато да могу да се повучем из истраживања у било ком тренутку, без образложења и последица.

У Крагујевцу _____.

Appendix F: Data Contribution and Consent form_Companies (in Serbian)

УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ

ИЗЈАВА О НАМЕРИ ИСТРАЖИВАЧА И ИЗЈАВА О ДОБРОВОЉНОМ УЧЕШЋУ ПРЕДСТАВНИКА КОМПАНИЈЕ У ИСТРАЖИВАЧКОМ ПРОЈЕКТУ

Назив истраживања: Investigating authentic forms of assessment in testing English for specific purpose speaking skills(Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику за посебне намене)

Истраживач: Милан Милановић, професор енглеског језика, докторанд

Поштованисараднице,

позвани сте да учествујете у пројекту под називом *Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику струке*, који има за циљ да утврди колико су аутентичне постојеће методе тестирања усмених језичких вештина на енглеском језику који се учи на Економском факултету у Крагујевцу. Пројекат, такође, има за циљ да утврди да ли су говорне вештине које студенти стичу учећи енглески језик на факултету управо оне вештине које су Вашој фирми потребне на тржишту рада.

Уколико пристанете да учествујете у истраживачком пројекту, Ваша помоћ подразумева следеће:

- кратак опис контекста у коме се у Вашој фирми користи говорни облик енглеског језика (уз пример конкретног језичког задатка који запослени треба да испуни)
- попуњавање упитника који је истраживач припремио, а који описује вештине говора на енглеском језику које сматрате пожељним код запослених сарадника економске струке (описи су на енглеском језику и имају облик “can do”).

Сви лични подаци, као и подаци везани за Вашу фирму су анонимни. Уколико одлучите да помогнете у спровођењу истраживања, а одлучите да се повучете из истог, слободни сте то да учините у било ком тренутку, без образложења и последица.

Ја, _____ (представник компаније _____), потписујем да добровољно пристајем да учествујем у овом истраживању. Истовремено потврђујем даразумем своју улогу и обавезе у реализацији истог. Осим тога, потврђујем да ми је

познато да могу да се повучем из истраживања у било ком тренутку, без образложења и последица.

У Крагујевцу _____.

Appendix G: Data Contribution Consent form_Companies (in English)

UNIVERSITY OF KRAGUJEVAC

INFORMATION STATEMENT AND DATA CONTRIBUTION CONSENT FORM FOR COMPANY REPRESENTATIVES

Research Title: Investigating authentic forms of assessment in testing English for specific purpose speaking skills (Испитивање аутентичних облика провере знања у тестирању говорних вештина на енглеском језику струке)

Researcher: Milan Milanović, English language instructor, doctoral student

Dear informant,

You are kindly invited to participate in the research entitled *Investigating authentic forms of assessment in testing English for specific purpose speaking skills*, the objective of which is to determine the level of authenticity of test tasks used to assess spoken English skills at the Faculty of Economics in Kragujevac. The research project aims at determining whether the English skills students develop at the Faculty correspond to the skills your Company requires in the labor market.

Should you give your consent to participate in this research, you will be asked to do the following:

- provide a brief description of the context in which spoken English is used in your Company (accompanied by a concrete example of the language task your employees are asked to complete)
- fill out the questionnaire prepared by the researcher, describing spoken English language skills that you consider preferable with your prospective employees who have educational background in economics (descriptions, in English language, take “can do...” form).

Also note that all personal and your Company details gathered in the course of the research are anonymous. If you decide to participate, you are free to withdraw from further participation in the research at any time without having to give a reason and without consequence.

I, _____ (the representative of _____ company), confirm by signing that I voluntarily agree to participate in this research. At the same time, I confirm that I understand my role and duties in the realization of the study. In addition, I confirm that I know that I can withdraw from further participation in the research at any time without having to give a reason and without consequence.

In Kragujevac, _____.

Appendix H: Context-based questionnaire: General Context

1	Your company performs business_____.	a) locally	b) internationally	c) both a and b
2	Your employees are required to use spoken English in business communication.	c) yes		d) no
3	If yes, what is the frequency of using English for business communication?	c) daily		d) occasionally
4	Your company employs business graduates majoring in one of the following: <i>Marketing, Management, Accounting and Business Finance.</i>	c) yes		d) no
5	Your company expects business graduates to be able to use oral English skills in business communication.	c) yes		d) no
6	When an individual speaks English, they apply ____ style(s).	d) conversational	e) presentational	f) both
7	Rank in the order of importance the following speaking tasks in English (1 being the most important, 7 being the least important):	informal conversation _____ phone call _____ group presentation _____ interview _____ giving a statement – formal (e.g. PR) _____ chat with colleagues _____ providing explanation/description (short monologue) _____		

Appendix I: Context-based questionnaire: Business Presentations

7	When they present in English, your employees are expected to do it_____.	d) individually	e) in a group	f) both a and b
8	When they present in English, individuals talk for_____min.	d) less than 5	e) 5-10	f) more than 10
9	When they present in English, the presentation can take place _____.	d) live	e) via video-conference call	f) both a and b
10	In an average business presentation, the number of the people in the audience is in the following range:	d) 1-5	e) 6-10	f) more than 10
11	People in the audience are:	d) colleagues	e) business associates/clients	f) both a and b
12	The communication and setting during presentations are:	d) formal	e) informal	f) both a and b
13	In an average business presentation, the people in the audience ask questions related to the content of the presentation.	c) yes		d) no
14	While presenting, the presenter(s) is/are required to manipulate equipment/use visuals/perform demonstrations.	c) yes		d) no
15	While presenting, the presenter(s) is expected to use technical words/specialized vocabulary? (e.g. related to the products/	c) yes		d) no

	production/specificities of the company itself, etc.)		
16	When presenting in English, your employees, with educational background in economics are expected to demonstrate the knowledge they gained in university.	c) yes	d) no
17	Can you rank the following in the order of importance (1 being the most important, 3 being the least important in a presentation)?	self-confidence and persuasiveness _____ clear organization and structure _____ native-like pronunciation _____	
18	Can you rank the following in the order of importance (1 being the most important, 3 being the least important)?	grammatical accuracy _____ fluency and voice projection _____ content and technical vocabulary _____	
19	Can you provide examples of presenting in English (consider who the presentations is delivered for? in what setting? how long was it? are there any special materials that presenters provided?)		
20	Can you provide any criteria by which you judge the success of a presentation in English?		

Appendix J: Assessor's rating scale (analytic): Rating scale: Oral presentation (group work)

	Above Expectations 4	Meets Expectations 3	Below Expectations 2	Poor 1	Total
Group Dynamics & Presentation Structure	Group stays within the time allotted: 3 SS (8-10 min), 4 SS (12-15 min); group members organized in movement and standing positions; Exceptional structure with intro, body, closing; smooth transitions; easy to follow.	Most members stay within time allotted: 3 SS (8-10 min), 4 SS (12-15 min); clear structure with intro, body, closing structure; most transitions appropriate; easy to follow.	Half of the members stay within time limit: 3 SS (8-10 min), 4 SS (12-15 min); Vague structure with intro, body, closing; transitions partially used; not easy to follow.	Few members stay within time limit 3 SS (8-10 min), 4 SS (12-15 min); disorganized structure; poor transitions; hard to follow.	/4
Visuals & Audience Engagement	PPP with graphic organizers provided. Visuals are relevant and interesting; all images illustrate the points/details of the presentation.	PPP with graphic organizers provided. Most visuals are relevant and interesting; most images illustrate the points/details of the presentation.	Irrelevant PPP and poorly designed graphic organizers. Very few visuals are relevant and interesting; hardly any of the images illustrate the points/details of the presentation.	Irrelevant PPP and poorly designed graphic organizers. Very few visuals are relevant and interesting; hardly any of the images illustrate the points/details of the presentation.	/4
Non-Verbal Comm.	Very expressive, confident, relaxed; appropriate posture and gestures	Mostly expressive, confident, relaxed; mostly appropriate posture and gestures	Somewhat expressive, confident, relaxed; Static posture and gestures	Not very expressive, confident, relaxed; Awkward posture and gestures	/4
Verbal Comm.	Project voice very well; very clear articulation; no hesitation; natural rhythm and pacing; use emphasis	Project voice; mostly clear articulation; little hesitation; natural rhythm and pacing; use emphasis	Do not project voice well enough; some clear articulation; a lot of hesitation; struggle to produce natural rhythm and pacing; little emphasis	Soft voice; mostly incomprehensible articulation; excessive hesitation; lack natural rhythm and pacing; lack emphasis, some use Serbian in the lack of English words	/4
Grammar and Vocabulary (group)	Excellent command of spoken grammar, with hardly noticeable mistakes. Wide range of topic-appropriate and <i>Marketing</i> -related vocabulary.	Very good grammar with a few mistakes that do not hinder meaning. Sufficient use of topic-related vocabulary. Uses some <i>Marketing</i> -related vocabulary items.	Inconsistent grammar with many mistakes. Repetitive vocabulary with few topic-appropriate items; occasional <i>Marketing</i> -related vocabulary items.	Grammar mistakes so numerous that meaning does not come through. Very basic, general vocabulary; no <i>Marketing</i> -related vocabulary items	/4
					/20:2 (max. 10 pts.)

Appendix K: Assessor's rating scale (holistic): Rating scale: Oral presentation (group work)

/10	Above Expectations 9-10	Meets Expectations 7-8	Below Expectations 5-6	Poor 3-4	Unsatisfactory 1-2
<p>Group Dynamics & Presentation Structure Visuals & Audience Engagement Non-Verbal Comm. Verbal Comm. Grammar and Vocabulary (group)</p>	<p>Group stays within the time allotted: 3 SS (8-10 min), 4 SS (12-15 min); group members organized; Exceptional presentation structure; smooth transitions; easy to follow. PPP with graphic organizers provided; visuals are relevant and interesting; very expressive, can project voice very well; can articulate well; persuasive; excellent command of spoken grammar, with hardly noticeable mistakes; wide range of topic-appropriate and <i>Marketing</i>-related vocabulary; excellent overall impression</p>	<p>Most members stay within time allotted: 3 SS (8-10 min), 4 SS (12-15 min); clear presentation; most transitions appropriate; easy to follow; PPP with graphic organizers provided; most visuals are relevant and interesting; mostly expressive; confident, relaxed; mostly appropriate posture and gestures; project voice; mostly clear articulation; use persuasive language; very good grammar with a few mistakes; sufficient use of topic-related and <i>Marketing</i>-related vocabulary items, very good overall impression</p>	<p>Half of the members stay within time limit: 3 SS (8-10 min), 4 SS (12-15 min); Vague structure; not easy to follow; irrelevant PPP and poorly designed graphic organizers; hardly any of the images illustrate the points/details of the presentation; static posture and gestures; do not project voice well enough; not very persuasive; inconsistent grammar with many mistakes; repetitive vocabulary with few topic-appropriate items; occasional <i>Marketing</i>-related vocabulary items; mediocre impression</p>	<p>Few members stay within time limit 3 SS (8-10 min), 4 SS (12-15 min); Disorganized structure; poor transitions; hard to follow. Irrelevant PPP and poorly designed graphic organizers. Awkward posture and gestures Soft voice; mostly incomprehensible articulation; excessive hesitation; some use Serbian in the lack of English words; grammar mistakes so numerous that meaning does not come through. Very basic, general vocabulary; no <i>Marketing</i>-related vocabulary items; poor overall impression</p>	<p>The presentation is extremely short, there is no attempt to present in an organized way; presenters did not make any visuals; presenters seem nervous, ill-prepared, confused as to how to proceed; they lack English skills, address the audience in Serbian; extremely poor impression</p>

Appendix L: Student self- / peer-assessment rating scale (analytic): Rating scale: Oral presentation (group work)

	Above Expectations 4	Meets Expectations 3	Below Expectations 2	Poor 1	Total
Group Dynamics & Presentation Structure	Group stays within the time specified in the instructions; group members appear as very organized; Excellent structure with intro, body, closing; smooth transitions; easy to follow.	The presentation is a bit shorter/longer than required but still interesting and informative; Clear structure with intro, body, closing structure; most transitions appropriate; easy to follow.	The presentation longer shorter/longer than required for no good reason. Structure not clear with component parts in the wrong order (e.g. body comes before the introduction); some transition words used wrongly; not easy to follow.	The presentation is too short/long; Disorganized structure; presenters not using transition words; very hard to follow.	/4
Visuals & Audience Engagement	PPP with graphic organizers provided. Visuals are relevant and interesting; all images illustrate the points/details of the presentation.	PPP with graphic organizers provided. Most visuals are relevant and interesting; most images illustrate the points/details of the presentation.	Acceptable PPP but with poorly designed graphic organizers. A few visuals are relevant and interesting; hardly any of the images illustrate the points/details of the presentation.	Irrelevant PPP and poorly designed graphic organizers. Very few visuals are relevant and interesting; hardly any of the images illustrate the points/details of the presentation.	/4
Non-Verbal Comm.	Presenters are excellent at expressing their ideas, everyone is confident, relaxed; appropriate posture and gestures	Mostly expressive, confident, relaxed; the posture and some gestures are sometimes awkward	Some presenters expressive, confident, relaxed, but most of them are not. Standing without movement or intention to employ non-verbal communication	Presenters very nervous and without confidence; everyone seems ill-prepared; inappropriate posture and gestures	/4
Verbal Comm.	Presenters speak loud and clear; they are easy to understand, they do not speak either fast or slowly and sound natural	Most of the group members speak clearly; a few times there is a pause in speech; they mostly sound natural with appropriate rhythm	Some presenters are hard to hear or understand; there is a lot of hesitation (e.g. “umm, er, hmm”); sound like they memorized the script	Very difficult to hear or understand, use Serbian when they cannot express themselves in English; many false starts and/or silence; very unnatural	/4
Grammar and Vocabulary (group)	Very difficult to notice any grammar mistakes. Many words coming from the units in English and <i>Marketing</i> courses.	Sometimes there is a grammar mistake, but it is not a serious one. I can recognize many words covered by the course in English (some in <i>Marketing</i> , too)	Grammar seems problematic. Many group members cannot apply grammar rules and the message is difficult to understand. Simple vocabulary with a few words covered by the course.	Very poor grammar, many mistakes making it impossible to understand what presenters are talking about. Very basic words used.	/4
					/20:2 (max. 10 pts.)

Appendix M: Student self- / peer-assessment rating scale (holistic): Rating scale: Oral presentation (group work)

/10	Above Expectations 9-10	Meets Expectations 7-8	Below Expectations 5-6	Poor 3-4	Unsatisfactory 1-2
<p>Group Dynamics & Presentation Structure Visuals & Audience Engagement Non-Verbal Comm. Verbal Comm. Grammar and Vocabulary</p>	<p>The group stays within the time allotted; group members are exceptionally well prepared with appropriate movements on the stage; visuals are quite relevant and interesting, the audience is engaged and lively; the presenters are confident and relaxed, speaking clearly and loudly without making grammar mistakes; many words coming from courses in English language and <i>Marketing</i></p>	<p>The presentation lasts a bit shorter/longer but for a good reason; presenters are standing or moving in a natural manner; the presentation is easy to follow and it includes excellent transitions; relevant and interesting visuals; the presenters use the body language in the appropriate manner; most presenters speak English clearly, sometimes they make pauses but it seem natural; the vocabulary words come from courses in English and <i>Marketing</i></p>	<p>The presentation is shorter/longer than it should be. It is not clear what the presentation is all about, some transition words well used; presenters seem confused, standing and moving awkwardly; the visuals are too simple/complicated and not interesting; presenters seem nervous; Some presenters are hard to hear or understand; there is a lot of hesitation (e.g. "umm, er, hmm"); sound like they memorized the script Grammar seems problematic. Simple vocabulary with a few words covered by the course.</p>	<p>The presentation is too short/long; no clear structure; Disorganized structure; no transition words; poorly designed visuals; presenters very nervous and without confidence; everyone seems ill-prepared; inappropriate posture and gestures Very difficult to hear or understand, use Serbian when they cannot express themselves in English; very unnatural Very poor grammar, many mistakes making it impossible to understand what presenters are talking about. Very basic words used.</p>	<p>The presentation is extremely short, no clear structure; PPP/visuals missing; students not prepared and nervous; no attempt to present in English; presenters use Serbian when addressing the audience</p>

Appendix N: Assessor's (and student self/ peer-rating) rating scale (holistic): Rating scale: Individual short presentation

5	Excellent	Can communicate ideas clearly and in a structured manner, providing appropriate examples; uses discourse markers and speaks coherently; pronunciation clear; minor grammar mistakes; topic-appropriate Business English vocabulary
4	Very good	Can provide a coherent account of the problem with matching examples; some discourse markers used out of place; minor slips of tongue, mostly General English vocabulary with a few items coming from Business English register; polite and professional
3	Good	Can state the problem and talk about it; speech interrupted with hesitation and false starts but mostly comprehensible; some vocabulary items mispronounced; one or two words coming from Business English register; polite
2	Poor	Ideas poorly organized and hesitation and many false starts; no examples; relies on the vocabulary provided in the prompt, nervous
0-1	Unsatisfactory	Shows little or no attempt to talk about the topic; repeats the prompt; addresses the interlocutor in Serbian; non-cooperative
/5x2 (max. 10pt)	Comments:	

Appendix O: Self-evaluation checklist - shuffled (spoken interaction and production in English)

A:

Read the descriptors below and tick (✓) ONLY THE BOX showing what you CAN do without help:	
SPOKEN INTERACTION	I can do this without help
I can get simple practical information (e.g., asking for directions, booking accommodation)	
I can take part in routine formal discussion on familiar subjects in my academic or professional field if it is conducted in clearly articulated speech in standard English	
I can greet other people and introduce myself	
I can handle numbers, quantities, cost and time	
I can exchange detailed factual information on matters within my academic or professional field	
I can exchange, check and confirm factual information on familiar routine and non-routine matters within my field with some confidence	
I can make and respond to invitations, suggestions, apologies and requests for permission	
I can account for and sustain my opinion in discussion by providing relevant explanations, arguments and comments	
I can say who I am, ask someone's name and introduce someone	
I can sustain an extended conversation or discussion on most topics that are familiar or of personal interest but may sometimes need help in communicating my thoughts	
I can handle most practical tasks in everyday situations (e.g., making telephone enquiries, asking for a refund, negotiating purchase)	
I can say I don't understand, ask people to repeat what they say or speak more slowly, attract attention and ask for help	
I can cope linguistically with potentially complex problems in routine situations (e.g., complaining about goods and services)	
I can express agreement and disagreement	
I can handle short social exchanges and make myself understood if people help me	
I can participate effectively in extended discussions and debates on subjects of personal, academic or professional interest, marking clearly the relationship between ideas	
I can ask people for things and give people things, saying "please" and "thank you" as appropriate	
I can express and respond to feelings and attitudes (e.g., surprise, happiness, sadness, interest, uncertainty, indifference)	
I can express, negotiate and respond sensitively to feelings, attitudes, opinions, tone, viewpoints	
I can participate in short conversations in routine contexts on topics of interest	
I can discuss current professional/learning targets in relation to future work or study options	
I can handle personal interviews with ease, taking initiatives and expanding ideas with little help from an interviewer	
I can take some initiatives in an interview/ consultation (e.g., bring up a new subject) but am very dependent on the interviewer to provide support	
I can make simple purchases, using pointing and gestures to support what I say	

I can obtain detailed information and can ask for and follow detailed directions	
I can say what I like or dislike	
I can provide concrete information required in an interview/consultation (e.g., describe symptoms to a doctor), but with limited precision	
I can cope adequately with emergencies (e.g., summon medical assistance, telephone the police or breakdown service)	
I can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow-up questions	
I can express my thoughts about abstract or cultural topics such as music or films, and give brief comments on the views of others	
I can explain why something is a problem, discuss what to do next, compare and contrast alternatives	
I can ask how someone is and say how I am	
I can reply in an interview to simple direct questions about personal details if these are spoken very slowly and clearly in standard English	
I can explain a problem to my teacher/manager/superior	
I can help along the progress of a project by inviting others to join in, express their opinions, etc.	
I can express what I feel in simple terms, and express thanks appropriately	
I can carry out an effective, fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies	
I can ask and answer simple direct questions on very familiar topics (e.g., family, student life, work) with help from the person I am talking to	
I can discuss what to do, where to go, make arrangements to meet (e.g., in the evening, at the weekend)	
I can ask and answer simple questions about familiar topics (e.g., weather, hobbies, social life, music, sport)	
I can ask and answer simple questions about things that have happened (e.g., yesterday, last week, last year)	
I can make simple transactions (e.g., in shops, post offices, railway stations) and order something to eat or drink	
I can handle simple telephone calls (e.g., say who is calling, ask to speak to someone, give my number)	

B:

Read the descriptors below and tick (✓) ONLY THE BOX showing what you CAN do without help:	
SPOKEN PRODUCTION	I can do this without help
I can say what I usually do at home, at school/college, at work, in my free time	
I can give detailed accounts of problems and incidents (e.g., reporting a theft, traffic accident)	
I can describe my qualifications and previous experience to an official	
I can spell my name and address	
I can give clear detailed descriptions on a wide range of subjects relating to my field, expanding and supporting ideas with subsidiary points and relevant examples	
I can deliver short rehearsed announcements and statements on everyday matters within my field	
I can explain a viewpoint on a topical issue, giving the advantages and disadvantages of various options	
I can give basic personal information about myself (e.g., name, age, address, family, subjects of study, job) using set phrases	
I can give a short and straightforward prepared presentation on a chosen topic in	

my academic or professional field in a reasonably clear and precise manner	
I can explain simply how to use a piece of equipment	
I can say the letters of the alphabet	
I can depart spontaneously from a prepared text and follow up points raised by an audience	
I can give short simple descriptions of events or tell a simple story	
I can pass on a simple message	
I can use simple words and phrases to describe where I live	
I can use simple words and phrases to describe people I know	
I can briefly give reasons and explanations for opinions, plans and actions	
I can give a clear, systematically developed presentation on a topic in my field, with highlighting of significant points and relevant supporting detail	
I can describe myself, my family and other people I know	
I can develop a clear coherent argument, linking ideas logically and expanding and supporting my points with appropriate examples	
I can describe personal experiences, reactions, dreams, hopes, ambitions, real, imagined or unexpected events	
I can make a very short rehearsed statement (e.g., to introduce a speaker)	
I can give a straightforward description of a subject within my academic or professional field, presenting it as a linear sequence of points	
I can outline an issue or a problem clearly, speculating about causes, consequences and hypothetical situations	
I can give a short rehearsed presentation on a familiar subject in my academic or professional field	
I can summarise short discursive or narrative material (e.g., written text, radio, television)	
I can give a simple summary of short written texts	
I can deliver announcements on most general topics with a degree of clarity, fluency and spontaneity which causes no strain or inconvenience to the listener	
I can describe my educational background and subjects of study	
I can narrate a story or relate the plot of a film or book	
I can develop an argument well enough to be followed without difficulty most of the time	
I can describe past activities and personal experiences (e.g., what I did at the weekend)	
I can give simple descriptions of things and make straightforward comparisons	
I can explain what I like and don't like about something	
I can deliver very short rehearsed announcements of predictable learnt content	

Appendix P: Self-evaluation checklist – ordered (with corresponding CEFR levels)

A:

SPOKEN INTERACTION	CEFR level
I can greet other people and introduce myself	A1.1SI
I can ask how someone is and say how I am	A1.2SI
I can say who I am, ask someone's name and introduce someone	A1.3SI
I can say I don't understand, ask people to repeat what they say or speak more slowly, attract attention and ask for help	A1.4SI
I can ask and answer simple direct questions on very familiar topics (e.g., family, student life, work) with help from the person I am talking to	A1.5SI
I can ask people for things and give people things, saying "please" and "thank you" as appropriate	A1.6SI
I can handle numbers, quantities, cost and time	A1.7SI
I can make simple purchases, using pointing and gestures to support what I say	A1.8SI
I can reply in an interview to simple direct questions about personal details if these are spoken very slowly and clearly in standard English	A1.9SI
I can ask and answer simple questions about things that have happened (e.g., yesterday, last week, last year)	A2.10SI
I can handle simple telephone calls (e.g., say who is calling, ask to speak to someone, give my number)	A2.11SI
I can make simple transactions (e.g., in shops, post offices, railway stations) and order something to eat or drink	A2.12SI
I can get simple practical information (e.g., asking for directions, booking accommodation)	A2.13SI
I can handle short social exchanges and make myself understood if people help me	A2.1SI
I can participate in short conversations in routine contexts on topics of interest	A2.2SI
I can make and respond to invitations, suggestions, apologies and requests for permission	A2.3SI
I can say what I like or dislike	A2.4SI
I can express agreement and disagreement	A2.5SI
I can explain a problem to my teacher/manager/superior	A2.6SI
I can express what I feel in simple terms, and express thanks appropriately	A2.7SI
I can discuss what to do, where to go, make arrangements to meet (e.g., in the evening, at the weekend)	A2.8SI
I can ask and answer simple questions about familiar topics (e.g., weather, hobbies, social life, music, sport)	A2.9SI
I can provide concrete information required in an interview/consultation (e.g., describe symptoms to a doctor), but with limited precision	B1.10SI
I can take some initiatives in an interview/consultation (e.g., bring up a new subject) but am very dependent on the interviewer to provide support	B1.11SI
I can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow-up questions	B1.12SI
I can sustain an extended conversation or discussion on most topics that are familiar or of personal interest but may sometimes need help in communicating my thoughts	B1.1SI
I can take part in routine formal discussion on familiar subjects in my academic or professional field if it is conducted in clearly articulated speech in	B1.2SI

standard English	
I can exchange, check and confirm factual information on familiar routine and non-routine matters within my field with some confidence	B1.3SI
I can express and respond to feelings and attitudes (e.g., surprise, happiness, sadness, interest, uncertainty, indifference)	B1.4SI
I can express my thoughts about abstract or cultural topics such as music or films, and give brief comments on the views of others	B1.5SI
I can explain why something is a problem, discuss what to do next, compare and contrast alternatives	B1.6SI
I can discuss current professional/learning targets in relation to future work or study options	B1.7SI
I can obtain detailed information and can ask for and follow detailed directions	B1.8SI
I can handle most practical tasks in everyday situations (e.g., making telephone enquiries, asking for a refund, negotiating purchase)	B1.9SI
I can carry out an effective, fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies	B2.10SI
I can participate effectively in extended discussions and debates on subjects of personal, academic or professional interest, marking clearly the relationship between ideas	B2.2SI
I can account for and sustain my opinion in discussion by providing relevant explanations, arguments and comments	B2.3SI
I can express, negotiate and respond sensitively to feelings, attitudes, opinions, tone, viewpoints	B2.4SI
I can exchange detailed factual information on matters within my academic or professional field	B2.5SI
I can help along the progress of a project by inviting others to join in, express their opinions, etc.	B2.6SI
I can cope linguistically with potentially complex problems in routine situations (e.g., complaining about goods and services)	B2.7SI
I can cope adequately with emergencies (e.g., summon medical assistance, telephone the police or breakdown service)	B2.8SI
I can handle personal interviews with ease, taking initiatives and expanding ideas with little help from an interviewer	B2.9SI

B:

SPOKEN PRODUCTION	CEFR level
I can say the letters of the alphabet	A1.1SP
I can spell my name and address	A1.2SP
I can give basic personal information about myself (e.g., name, age, address, family, subjects of study, job) using set phrases	A1.3SP
I can pass on a simple message	A1.4SP
I can use simple words and phrases to describe where I live	A1.5SP
I can use simple words and phrases to describe people I know	A1.6SP
I can make a very short rehearsed statement (e.g., to introduce a speaker)	A1.7SP
I can give a short rehearsed presentation on a familiar subject in my academic or professional field	A2.10SP
I can describe myself, my family and other people I know	A2.1SP
I can describe my educational background and subjects of study	A2.2SP
I can say what I usually do at home, at school/college, at work, in my free time	A2.3SP
I can describe my qualifications and previous experience to an official	A2.4SP
I can give short simple descriptions of events or tell a simple story	A2.5SP
I can describe past activities and personal experiences (e.g., what I did at the weekend)	A2.6SP

I can explain what I like and don't like about something	A2.7SP
I can give simple descriptions of things and make straightforward comparisons	A2.8SP
I can deliver very short rehearsed announcements of predictable learnt content	A2.9SP
I can explain simply how to use a piece of equipment	B1.10SP
I can give a straightforward description of a subject within my academic or professional field, presenting it as a linear sequence of points	B1.1SP
I can narrate a story or relate the plot of a film or book	B1.2SP
I can describe personal experiences, reactions, dreams, hopes, ambitions, real, imagined or unexpected events	B1.3SP
I can briefly give reasons and explanations for opinions, plans and actions	B1.4SP
I can develop an argument well enough to be followed without difficulty most of the time	B1.5SP
I can give a simple summary of short written texts	B1.6SP
I can give detailed accounts of problems and incidents (e.g., reporting a theft, traffic accident)	B1.7SP
I can deliver short rehearsed announcements and statements on everyday matters within my field	B1.8SP
I can give a short and straightforward prepared presentation on a chosen topic in my academic or professional field in a reasonably clear and precise manner	B1.9SP
I can give clear detailed descriptions on a wide range of subjects relating to my field, expanding and supporting ideas with subsidiary points and relevant examples	B2.1SP
I can explain a viewpoint on a topical issue, giving the advantages and disadvantages of various options	B2.2SP
I can develop a clear coherent argument, linking ideas logically and expanding and supporting my points with appropriate examples	B2.3SP
I can outline an issue or a problem clearly, speculating about causes, consequences and hypothetical situations	B2.4SP
I can summarise short discursive or narrative material (e.g., written text, radio, television)	B2.5SP
I can deliver announcements on most general topics with a degree of clarity, fluency and spontaneity which causes no strain or inconvenience to the listener	B2.6SP
I can give a clear, systematically developed presentation on a topic in my field, with highlighting of significant points and relevant supporting detail	B2.7SP
I can depart spontaneously from a prepared text and follow up points raised by an audience	B2.8SP

Appendix Q: Can-do evaluation checklist - shuffled (spoken interaction and production in English for subject specialist informants)

A:

Read the descriptors below and tick (√) ONLY THE BOX showing what you think your prospective employees should know how to do:	
SPOKEN INTERACTION	can do this without help
can get simple practical information (e.g., asking for directions, booking accommodation)	
can take part in routine formal discussion on familiar subjects in my academic or professional field if it is conducted in clearly articulated speech in standard English	
can greet other people and introduce myself	
can handle numbers, quantities, cost and time	
can exchange detailed factual information on matters within my academic or professional field	
can exchange, check and confirm factual information on familiar routine and non-routine matters within my field with some confidence	
can make and respond to invitations, suggestions, apologies and requests for permission	
can account for and sustain my opinion in discussion by providing relevant explanations, arguments and comments	
can say who I am, ask someone's name and introduce someone	
can sustain an extended conversation or discussion on most topics that are familiar or of personal interest but may sometimes need help in communicating my thoughts	
can handle most practical tasks in everyday situations (e.g., making telephone enquiries, asking for a refund, negotiating purchase)	
can say I don't understand, ask people to repeat what they say or speak more slowly, attract attention and ask for help	
can cope linguistically with potentially complex problems in routine situations (e.g., complaining about goods and services)	
can express agreement and disagreement	
can handle short social exchanges and make myself understood if people help me	
can participate effectively in extended discussions and debates on subjects of personal, academic or professional interest, marking clearly the relationship between ideas	
can ask people for things and give people things, saying "please" and "thank you" as appropriate	
can express and respond to feelings and attitudes (e.g., surprise, happiness, sadness, interest, uncertainty, indifference)	
can express, negotiate and respond sensitively to feelings, attitudes, opinions, tone, viewpoints	
can participate in short conversations in routine contexts on topics of interest	
can discuss current professional/learning targets in relation to future work or study options	
can handle personal interviews with ease, taking initiatives and expanding	

ideas with little help from an interviewer	
can take some initiatives in an interview/ consultation (e.g., bring up a new subject) but am very dependent on the interviewer to provide support	
can make simple purchases, using pointing and gestures to support what I say	
can obtain detailed information and can ask for and follow detailed directions	
can say what I like or dislike	
can provide concrete information required in an interview/consultation (e.g., describe symptoms to a doctor), but with limited precision	
can cope adequately with emergencies (e.g., summon medical assistance, telephone the police or breakdown service)	
can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow-up questions	
can express my thoughts about abstract or cultural topics such as music or films, and give brief comments on the views of others	
can explain why something is a problem, discuss what to do next, compare and contrast alternatives	
can ask how someone is and say how I am	
can reply in an interview to simple direct questions about personal details if these are spoken very slowly and clearly in standard English	
can explain a problem to my teacher/manager/superior	
can help along the progress of a project by inviting others to join in, express their opinions, etc.	
can express what I feel in simple terms, and express thanks appropriately	
can carry out an effective, fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies	
can ask and answer simple direct questions on very familiar topics (e.g., family, student life, work) with help from the person I am talking to	
can discuss what to do, where to go, make arrangements to meet (e.g., in the evening, at the weekend)	
can ask and answer simple questions about familiar topics (e.g., weather, hobbies, social life, music, sport)	
can ask and answer simple questions about things that have happened (e.g., yesterday, last week, last year)	
can make simple transactions (e.g., in shops, post offices, railway stations) and order something to eat or drink	
can handle simple telephone calls (e.g., say who is calling, ask to speak to someone, give my number)	

B:

Read the descriptors below and tick (✓) ONLY THE BOX showing what you think your prospective employees should know how to do:	
SPOKEN PRODUCTION	can do this without help
can say what I usually do at home, at school/college, at work, in my free time	
can give detailed accounts of problems and incidents (e.g., reporting a theft, traffic accident)	
can describe my qualifications and previous experience to an official	
can spell my name and address	
can give clear detailed descriptions on a wide range of subjects relating to my field, expanding and supporting ideas with subsidiary points and relevant examples	
can deliver short rehearsed announcements and statements on everyday matters within my field	
can explain a viewpoint on a topical issue, giving the advantages and disadvantages of various options	
can give basic personal information about myself (e.g., name, age, address, family, subjects of study, job) using set phrases	
can give a short and straightforward prepared presentation on a chosen topic in my academic or professional field in a reasonably clear and precise manner	
can explain simply how to use a piece of equipment	
can say the letters of the alphabet	
can depart spontaneously from a prepared text and follow up points raised by an audience	
can give short simple descriptions of events or tell a simple story	
can pass on a simple message	
can use simple words and phrases to describe where I live	
can use simple words and phrases to describe people I know	
can briefly give reasons and explanations for opinions, plans and actions	
can give a clear, systematically developed presentation on a topic in my field, with highlighting of significant points and relevant supporting detail	
can describe myself, my family and other people I know	
can develop a clear coherent argument, linking ideas logically and expanding and supporting my points with appropriate examples	
can describe personal experiences, reactions, dreams, hopes, ambitions, real, imagined or unexpected events	
can make a very short rehearsed statement (e.g., to introduce a speaker)	
can give a straightforward description of a subject within my academic or professional field, presenting it as a linear sequence of points	
can outline an issue or a problem clearly, speculating about causes, consequences and hypothetical situations	
can give a short rehearsed presentation on a familiar subject in my academic or professional field	
can summarise short discursive or narrative material (e.g., written text, radio, television)	
can give a simple summary of short written texts	
can deliver announcements on most general topics with a degree of clarity, fluency and spontaneity which causes no strain or inconvenience to the listener	
can describe my educational background and subjects of study	
can narrate a story or relate the plot of a film or book	

can develop an argument well enough to be followed without difficulty most of the time	
can describe past activities and personal experiences (e.g., what I did at the weekend)	
can give simple descriptions of things and make straightforward comparisons	
can explain what I like and don't like about something	
can deliver very short rehearsed announcements of predictable learnt content	

Appendix R: Self-evaluation checklist – ordered (spoken interaction and production in English for subject specialist informants)

A:

SPOKEN INTERACTION	CEFR level
can greet other people and introduce myself	A1.1SI
can ask how someone is and say how I am	A1.2SI
can say who I am, ask someone's name and introduce someone	A1.3SI
can say I don't understand, ask people to repeat what they say or speak more slowly, attract attention and ask for help	A1.4SI
can ask and answer simple direct questions on very familiar topics (e.g., family, student life, work) with help from the person I am talking to	A1.5SI
can ask people for things and give people things, saying "please" and "thank you" as appropriate	A1.6SI
can handle numbers, quantities, cost and time	A1.7SI
can make simple purchases, using pointing and gestures to support what I say	A1.8SI
can reply in an interview to simple direct questions about personal details if these are spoken very slowly and clearly in standard English	A1.9SI
can ask and answer simple questions about things that have happened (e.g., yesterday, last week, last year)	A2.10SI
can handle simple telephone calls (e.g., say who is calling, ask to speak to someone, give my number)	A2.11SI
can make simple transactions (e.g., in shops, post offices, railway stations) and order something to eat or drink	A2.12SI
can get simple practical information (e.g., asking for directions, booking accommodation)	A2.13SI
can handle short social exchanges and make myself understood if people help me	A2.1SI
can participate in short conversations in routine contexts on topics of interest	A2.2SI
can make and respond to invitations, suggestions, apologies and requests for permission	A2.3SI
can say what I like or dislike	A2.4SI
can express agreement and disagreement	A2.5SI
can explain a problem to my teacher/manager/superior	A2.6SI
can express what I feel in simple terms, and express thanks appropriately	A2.7SI
can discuss what to do, where to go, make arrangements to meet (e.g., in the evening, at the weekend)	A2.8SI
can ask and answer simple questions about familiar topics (e.g., weather, hobbies, social life, music, sport)	A2.9SI
can provide concrete information required in an interview/consultation (e.g., describe symptoms to a doctor), but with limited precision	B1.10SI
can take some initiatives in an interview/consultation (e.g., bring up a new subject) but am very dependent on the interviewer to provide support	B1.11SI
can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow-up questions	B1.12SI
can sustain an extended conversation or discussion on most topics that are familiar or of personal interest but may sometimes need help in communicating my thoughts	B1.1SI
can take part in routine formal discussion on familiar subjects in my academic or professional field if it is conducted in clearly articulated speech in standard English	B1.2SI
can exchange, check and confirm factual information on familiar routine and	B1.3SI

non-routine matters within my field with some confidence	
can express and respond to feelings and attitudes (e.g., surprise, happiness, sadness, interest, uncertainty, indifference)	B1.4SI
can express my thoughts about abstract or cultural topics such as music or films, and give brief comments on the views of others	B1.5SI
can explain why something is a problem, discuss what to do next, compare and contrast alternatives	B1.6SI
can discuss current professional/learning targets in relation to future work or study options	B1.7SI
can obtain detailed information and can ask for and follow detailed directions	B1.8SI
can handle most practical tasks in everyday situations (e.g., making telephone enquiries, asking for a refund, negotiating purchase)	B1.9SI
can carry out an effective, fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies	B2.10SI
can participate effectively in extended discussions and debates on subjects of personal, academic or professional interest, marking clearly the relationship between ideas	B2.2SI
can account for and sustain my opinion in discussion by providing relevant explanations, arguments and comments	B2.3SI
can express, negotiate and respond sensitively to feelings, attitudes, opinions, tone, viewpoints	B2.4SI
can exchange detailed factual information on matters within my academic or professional field	B2.5SI
can help along the progress of a project by inviting others to join in, express their opinions, etc.	B2.6SI
can cope linguistically with potentially complex problems in routine situations (e.g., complaining about goods and services)	B2.7SI
can cope adequately with emergencies (e.g., summon medical assistance, telephone the police or breakdown service)	B2.8SI
can handle personal interviews with ease, taking initiatives and expanding ideas with little help from an interviewer	B2.9SI

B:

SPOKEN PRODUCTION	CEFR level
can say the letters of the alphabet	A1.1SP
can spell my name and address	A1.2SP
can give basic personal information about myself (e.g., name, age, address, family, subjects of study, job) using set phrases	A1.3SP
can pass on a simple message	A1.4SP
can use simple words and phrases to describe where I live	A1.5SP
can use simple words and phrases to describe people I know	A1.6SP
can make a very short rehearsed statement (e.g., to introduce a speaker)	A1.7SP
can give a short rehearsed presentation on a familiar subject in my academic or professional field	A2.10SP
can describe myself, my family and other people I know	A2.1SP
can describe my educational background and subjects of study	A2.2SP
can say what I usually do at home, at school/college, at work, in my free time	A2.3SP
can describe my qualifications and previous experience to an official	A2.4SP
can give short simple descriptions of events or tell a simple story	A2.5SP
can describe past activities and personal experiences (e.g., what I did at the weekend)	A2.6SP
can explain what I like and don't like about something	A2.7SP
can give simple descriptions of things and make straightforward	A2.8SP

comparisons	
can deliver very short rehearsed announcements of predictable learnt content	A2.9SP
can explain simply how to use a piece of equipment	B1.10SP
can give a straightforward description of a subject within my academic or professional field, presenting it as a linear sequence of points	B1.1SP
can narrate a story or relate the plot of a film or book	B1.2SP
can describe personal experiences, reactions, dreams, hopes, ambitions, real, imagined or unexpected events	B1.3SP
can briefly give reasons and explanations for opinions, plans and actions	B1.4SP
can develop an argument well enough to be followed without difficulty most of the time	B1.5SP
can give a simple summary of short written texts	B1.6SP
can give detailed accounts of problems and incidents (e.g., reporting a theft, traffic accident)	B1.7SP
can deliver short rehearsed announcements and statements on everyday matters within my field	B1.8SP
can give a short and straightforward prepared presentation on a chosen topic in my academic or professional field in a reasonably clear and precise manner	B1.9SP
can give clear detailed descriptions on a wide range of subjects relating to my field, expanding and supporting ideas with subsidiary points and relevant examples	B2.1SP
can explain a viewpoint on a topical issue, giving the advantages and disadvantages of various options	B2.2SP
can develop a clear coherent argument, linking ideas logically and expanding and supporting my points with appropriate examples	B2.3SP
can outline an issue or a problem clearly, speculating about causes, consequences and hypothetical situations	B2.4SP
can summarise short discursive or narrative material (e.g., written text, radio, television)	B2.5SP
can deliver announcements on most general topics with a degree of clarity, fluency and spontaneity which causes no strain or inconvenience to the listener	B2.6SP
can give a clear, systematically developed presentation on a topic in my field, with highlighting of significant points and relevant supporting detail	B2.7SP
can depart spontaneously from a prepared text and follow up points raised by an audience	B2.8SP

Appendix S: Student attitudes questionnaire (A in English, B – in Serbian)

A: Note: Please tick only ONE answer which best describes your opinion.

		I totally disagree	I mostly disagree	I have no opinion	I mostly agree	I totally agree
1.	I study English so that I can communicate with foreigners.					
2.	I study English because I want to get a job with an international company.					
3.	The tasks we were solving this semester in English language 2 classes will help me outside classroom as well.					
4.	The presentation tasks helped me build my confidence when speaking in English					
5.	It is important for me to know the criteria based on which my performance is judged by the instructor.					
6.	I like the idea of judging my own performance by the same criteria the instructor uses to judge it.					
7.	At my future workplace I will need the skills of reading and listening more than any other English language skills.					
8.	I think that presentation skills will help me in my future career.					
9.	I think that English language should be taught throughout undergraduate studies.					
10.	It is easier for me to speak than to write in English.					
11.	I like tasks allowing me to choose how to solve them, e.g. by choosing a topic or preparation material for my presentation.					
12.	It is easier for me to write than to speak in English.					
13.	English is best learnt in a small group of students.					
14.	At my future workplace I will need the skills of writing and speaking more than any other English language skills.					
15.	It is easier for me to read than to listen to speech in English.					
16.	I feel more confident at speaking in English after delivering my oral presentation in this language.					
17.	It is easier for me to listen than to read in English.					
18.	The presentation tasks helped me build my confidence when speaking in English.					
19.	I like tasks resembling a project or tasks requiring group work.					
20.	I like the idea of judging my peers' performance by the same criteria I use to judge my own performance.					

B: Uputstvo: Štiklirajte samo JEDAN odgovor koji najbolje opisuje Vaše mišljenje.

		Uopšte se ne slažem	uglavnom se ne slažem	Nemam mišljenje	uglavnom se slažem	potpuno se slažem
1.	Engleski jezik učim da bih mogao/mogla da komuniciram sa strancima .					
2.	Engleski jezik učim jer želim da dobijem posao u firmi koja posluje sa inostranstvom.					
3.	Zadaci sa kojima smo se susretali tokom ovog semestra na predmetu Engleski jezik 2 će mi pomoći u budućnosti i van učionice.					
4.	Povratna informacija koju sam dobio/la po završenoj prezentaciji pomogla mi je da ispravim greške.					
5.	Važno mi je da znam na osnovu kojih kriterijuma me ocenjuje nastavnik.					
6.	Dopada mi se mogućnost da ocenjujem sebe na osnovu kriterijuma na osnovu kojih ocenjuje nastavnik.					
7.	U poslu će mi najviše trebati veštine slušanja i čitanja na engleskom jeziku.					
8.	Smatram da će mi veštine prezentovanja na engleskom jeziku pomoći u budućoj karijeri.					
9.	Smatram da engleski treba da se uči tokom sve 4 godine studija.					
10.	Lakše mi je da govorim na engleskom nego da pišem.					
11.	Dopadaju mi se zadaci u kojima mogu da biram kako ću da ih rešim. Npr. da samostalno biram temu i materijal za pripremu prezentacije.					
12.	Lakše mi je da pišem na engleskom nego da govorim.					
13.	Engleski jezik se bolje uči u manjoj grupi studenata.					
14.	U poslu će mi najviše trebati veštine govora i pisanja na engleskom jeziku.					
15.	Lakše mi je da čitam na engleskom jeziku nego da slušam.					
16.	Nakon usmene prezentacije na engleskom jeziku imam više samopouzdanja da govorim na ovom jeziku.					
17.	Lakše mi je da slušam govor na engleskom jeziku nego da čitam.					
18.	Zadaci sa prezentacijama su mi pomogli da steknem samopouzdanje kada govorim na engleskom jeziku.					
19.	Dopadaju mi se zadaci na engleskom jeziku koji liče na projekat ili na zadatak koji rešavam sa drugim članovima grupe.					
20.	Dopada mi se mogućnost da ocenjujem kolege na osnovu kriterijuma na osnovu kojih ocenjujem sebe.					

Appendix T: Self-evaluation checklist - shuffled (target; with or without help)

A:

Read the descriptors below and tick (✓) ONLY ONE box showing what your target is, or what you actually CAN do with or without help:			
SPOKEN INTERACTION	This is my target	I can now do this with help	I can now do this without help
I can get simple practical information (e.g., asking for directions, booking accommodation)			
I can take part in routine formal discussion on familiar subjects in my academic or professional field if it is conducted in clearly articulated speech in standard English			
I can greet other people and introduce myself			
I can handle numbers, quantities, cost and time			
I can exchange detailed factual information on matters within my academic or professional field			
I can exchange, check and confirm factual information on familiar routine and non-routine matters within my field with some confidence			
I can make and respond to invitations, suggestions, apologies and requests for permission			
I can account for and sustain my opinion in discussion by providing relevant explanations, arguments and comments			
I can say who I am, ask someone's name and introduce someone			
I can sustain an extended conversation or discussion on most topics that are familiar or of personal interest but may sometimes need help in communicating my thoughts			
I can handle most practical tasks in everyday situations (e.g., making telephone enquiries, asking for a refund, negotiating purchase)			
I can say I don't understand, ask people to repeat what they say or speak more slowly, attract attention and ask for help			
I can cope linguistically with potentially complex problems in routine situations (e.g., complaining about goods and services)			
I can express agreement and disagreement			
I can handle short social exchanges and make myself understood if people help me			
I can participate effectively in extended discussions and debates on subjects of personal, academic or professional interest, marking clearly the relationship between ideas			
I can ask people for things and give people things, saying "please" and "thank you" as appropriate			
I can express and respond to feelings and attitudes (e.g., surprise, happiness, sadness, interest, uncertainty, indifference)			
I can express, negotiate and respond sensitively to feelings,			

attitudes, opinions, tone, viewpoints			
I can participate in short conversations in routine contexts on topics of interest			
I can discuss current professional/learning targets in relation to future work or study options			
I can handle personal interviews with ease, taking initiatives and expanding ideas with little help from an interviewer			
I can take some initiatives in an interview/ consultation (e.g., bring up a new subject) but am very dependent on the interviewer to provide support			
I can make simple purchases, using pointing and gestures to support what I say			
I can obtain detailed information and can ask for and follow detailed directions			
I can say what I like or dislike			
I can provide concrete information required in an interview/consultation (e.g., describe symptoms to a doctor), but with limited precision			
I can cope adequately with emergencies (e.g., summon medical assistance, telephone the police or breakdown service)			
I can use a prepared questionnaire to carry out a structured interview, with some spontaneous follow-up questions			
I can express my thoughts about abstract or cultural topics such as music or films, and give brief comments on the views of others			
I can explain why something is a problem, discuss what to do next, compare and contrast alternatives			
I can ask how someone is and say how I am			
I can reply in an interview to simple direct questions about personal details if these are spoken very slowly and clearly in standard English			
I can explain a problem to my teacher/manager/superior			
I can help along the progress of a project by inviting others to join in, express their opinions, etc.			
I can express what I feel in simple terms, and express thanks appropriately			
I can carry out an effective, fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies			
I can ask and answer simple direct questions on very familiar topics (e.g., family, student life, work) with help from the person I am talking to			
I can discuss what to do, where to go, make arrangements to meet (e.g., in the evening, at the weekend)			
I can ask and answer simple questions about familiar topics (e.g., weather, hobbies, social life, music, sport)			
I can ask and answer simple questions about things that have happened (e.g., yesterday, last week, last year)			
I can make simple transactions (e.g., in shops, post offices, railway stations) and order something to eat or drink			
I can handle simple telephone calls (e.g., say who is calling, ask to speak to someone, give my number)			

B:

Read the descriptors below and tick (√) ONLY ONE box showing what your target is, or what you actually CAN do with or without help:			
SPOKEN PRODUCTION	This is my target	I can now do this with help	I can now do this without help
I can say what I usually do at home, at school/college, at work, in my free time			
I can give detailed accounts of problems and incidents (e.g., reporting a theft, traffic accident)			
I can describe my qualifications and previous experience to an official			
I can spell my name and address			
I can give clear detailed descriptions on a wide range of subjects relating to my field, expanding and supporting ideas with subsidiary points and relevant examples			
I can deliver short rehearsed announcements and statements on everyday matters within my field			
I can explain a viewpoint on a topical issue, giving the advantages and disadvantages of various options			
I can give basic personal information about myself (e.g., name, age, address, family, subjects of study, job) using set phrases			
I can give a short and straightforward prepared presentation on a chosen topic in my academic or professional field in a reasonably clear and precise manner			
I can explain simply how to use a piece of equipment			
I can say the letters of the alphabet			
I can depart spontaneously from a prepared text and follow up points raised by an audience			
I can give short simple descriptions of events or tell a simple story			
I can pass on a simple message			
I can use simple words and phrases to describe where I live			
I can use simple words and phrases to describe people I know			
I can briefly give reasons and explanations for opinions, plans and actions			
I can give a clear, systematically developed presentation on a topic in my field, with highlighting of significant points and relevant supporting detail			
I can describe myself, my family and other people I know			
I can develop a clear coherent argument, linking ideas logically and expanding and supporting my points with appropriate examples			
I can describe personal experiences, reactions, dreams, hopes, ambitions, real, imagined or unexpected events			
I can make a very short rehearsed statement (e.g., to introduce a speaker)			
I can give a straightforward description of a subject within my academic or professional field, presenting it as a linear sequence of points			
I can outline an issue or a problem clearly, speculating about causes, consequences and hypothetical situations			
I can give a short rehearsed presentation on a familiar subject in my academic or professional field			
I can summarise short discursive or narrative material (e.g., written text, radio, television)			
I can give a simple summary of short written texts			

I can deliver announcements on most general topics with a degree of clarity, fluency and spontaneity which causes no strain or inconvenience to the listener			
I can describe my educational background and subjects of study			
I can narrate a story or relate the plot of a film or book			
I can develop an argument well enough to be followed without difficulty most of the time			
I can describe past activities and personal experiences (e.g., what I did at the weekend)			
I can give simple descriptions of things and make straightforward comparisons			
I can explain what I like and don't like about something			
I can deliver very short rehearsed announcements of predictable learnt content			

About the author (biography)

Milan Milanović was born on 18 October, 1977 in Kragujevac, where he attended primary and secondary school. He enrolled in undergraduate program at the Faculty of Philology in Belgrade in 1996, and graduated in 2001 with GPA of 9.0 (9/10). In 2009, he enrolled in MA graduate studies and received his MA degree in 2010, with the GPA of 9.33. He started his postgraduate studies in 2011, when he enrolled in a doctoral program – Language module - at the Faculty of Philology, University of Belgrade.

He started his teaching career at the Faculty of Philology and Arts, University of Kragujevac, in 2003. In 2008, he was promoted to the position of a Senior Instructor of English language, teaching English language skills at the English Department. The following year, he expanded his professional expertise to ESP, when he started teaching Business English at the Faculty of Economics, University of Kragujevac. In 2015, he moved to Kuwait, where he works as an English Language Instructor and a Team Leader at the American University of the Middle East.

О аутору (биографија)

Милан Милановић рођен је 18.10.1977. у Крагујевцу, где је завршио основну и средњу школу. Основне студије уписао је на Филолошком факултету у Београду, 1996 године, а дипломирао је 2001. године са просечном оценом 9.0. Мастер академске студије уписао је 2009. године, а дипломирао је 2010. са просечном оценом 9.33. Постдипломске студије наставио је 2011. године, уписавши се на модул *Језик*, у оквиру докторских академских студија Филолошког факултета Универзитета у Београду.

Каријеру предавача започео је 2003. године, запосливши се на Филолошко-уметничком факултету Универзитета у Крагујевцу. 2008. године стекао је академско звање Вишег предавача за енглески језик. Наредне, 2009. године, почео је да предаје пословни енглески језик на Економском факултету Универзитета у Крагујевцу. Крајем 2015. године, преселио се у Кувајт, где и сада борава и ради као предавач и тим лидер

за енглески језик на високошколској институцији која носи назив „American University of the Middle East“.

Прилог 1.

Изјава о ауторству

Потписани Милан Милановић

Број уписа 11057д

Изјављујем


да је докторска дисертација под насловом

„Investigating Authentic Forms of Assessment in Testing English for Specific Purpose Speaking Skills”

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени, и
- да нисам кршио ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 26.08.2019.



Прилог 2.

**Изјава о истоветности штампане и електронске верзије
докторског рада**

Име и презиме аутора: Милан Милановић

Број уписа: 11057д

Студијски програм: Језик

Наслов рада: „Investigating Authentic Forms of Assessment in Testing English for Specific Purpose Speaking Skills”

Ментор: проф. др Оливера Дурбаба, редовни професор, Филолошки факултет
Универзитета у Београду

Потписани Милан Милановић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним станицама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 26.08.2019.



Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

„Investigating Authentic Forms of Assessment in Testing English for Specific Purpose Speaking Skills”,

која је моје ауторско дело.

Дисертацију са свим прилозима предао сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио.

1. Ауторство
2. Ауторство – некомерцијално
3. Ауторство - некомерцијално – без прераде
4. Ауторство - некомерцијално – делити под истим условима
5. Ауторство – без прераде
6. Ауторство - делити под истим условима

Потпис докторанда

У Београду, 26.08.2019.



