Relating Latvian Year 12 Examination in English to the CEFR

Contents

Introduction

The present paper describes the process of linking Latvian Year 12 Examination in English to the Common European Framework of Reference (CEFR) levels undertaken by a group of second year MA programme students in English Philology at the faculty of Humanities. The project was led by State Education Centre in 2010. The aim of the project was to define whether the examination complies with the requirements of the Common European Framework of Reference and can be used to assess school-leavers language proficiency according to the proficiency levels described in the CEFR.

The need for relation was defined by the introduction of new Education Standarts of Secondary Education in 2008. According to the new regulations, students are expected to reach levels B2-C1 if they study English as the first of second foreign language and level B1 if English is their third foreign language. Thus, it is a question of utmost importance to see whether the examination in its present form is capable of providing tasks for the given levels. Current publication presents the psychometric characteristics of the test and the main qualitative findings of the research exercise linking the Year 12 examination in English to Common European Framework of Reference levels.

Psychometric Characteristics of Latvian Year 12 Exam in English 2010 (Tatjana Kunda)

The aim of the quantitative analysis of Latvian Year 12 Examination in English 2010 is to define its psychometric characteristics. Quantitative analysis is an indispensable part of a test validation process. As Bachman (2004:3) puts it, 'an important kind of evidence that we collect to support test use is that which we derive from quantitative data [...] and the appropriate statistical analyses of these data.' Thus, statistical analysis can help make additional inferences about the quality of individual test items and tests on the whole.

Materials

The following materials were submitted for the study:

- Year 12 Examination in English descriptive statistics published by State Education Centre available at http://visc.gov.lv/eksameni/vispizgl/statistika.shtml;
- 22,638 students' scores in the listening, reading and language use parts of Year 12 Examination in English 2010.

Two types of software were used during the analysis:

- ITEMAN for Windows, Version 3.50, Copyright © 1995 by Assessment Systems Corporation to perform classical item analysis;
- Winsteps ® Rasch Measurement, Version 3.70.0.5, Copyright © 2009 John M. Linacre to perform Rasch analysis.

Methods

In order to ensure maximum efficiency of the research two types of statistical analysis were applied:

- procedures following the principles of Classical Test Theory (CTT);
- Rasch analysis (one-parameter Item Response Theory Model).

CTT was used in order to carry out initial analysis of the difficulty and discriminating ability of individual test items as well as test reliability. Even though CTT receives a certain amount of criticism for a number of limitations (e.g. dependency on a particular group, see Bachman, 2004:139), its procedures can still provide an insight into the quality of items. According to the pilot version of the Manual (2003:105), 'running a standard CTT analysis as the first step in empirical validation is always to be recommended', since it will provide 'a useful insight into the quality of individual items, and an indication of the reliability of the test as a whole.'

Cronbach's Alpha was considered in order to define the reliability of the exam parts. The following interpretation of the alpha was used:

- < 0.60 – unacceptable;
- 0.60 – 0.65 – undesirable;
- 0.65 – 0.70 – minimally acceptable;
- 0.70 – 0.80 – respectable;
- 0.80 – 0.90 – very good;
- > 0.90 – consider shortening the scale (Everitt 2006:108)

According to the Manual (2009:94), item difficulty level (p-value) and item discrimination are the first psychometric aspects when discussing appropriateness and usefulness of the test. Thus, classical item analysis was used to analyse the difficulty level (p-value) and discrimination index (D) of each individual item in the exam parts.

Item difficulty is the 'proportion of test takers who answered the item correctly (Bachman, 2004:122).' The following interpretation of item difficulty index was applied (ibid.):

- > 0.85 – the item is too easy;
- 0.3 – 0.8 – acceptable item difficulty;
- < 0.25 – the item is too difficult.

In a multiple-choice test effective distracters are expected to the value of difficulty index of at least 0.10. For a norm-referenced test the most informative are the items whose difficulty index is around 0.50. Such items would provide higher discrimination and, thus, provide a better spread of the scores, which is essential for a more accurate measurement of test-takers' proficiency.

Item discrimination is the extent to which the item discriminates between different groups of test takers.'(Bachman, 2004:122). Bachman also suggests that for a norm-referenced test, whose purpose is to 'differentiate among test takers at different levels of ability (ibid, 30), the desirable value of discrimination indices is above 0.30. At the same time Ebel and Frisbie (1991, in Oermann and Gaberson, 1998:155) suggest the following interpretation of discrimination index, which was also used in the present research:

- > 0.40 – very good items;
- 0.30 – 0.39 – reasonably good items;
- 0.20 – 0.29 – the item needs to be improved;

- < 0.19 – should be revised or not used again.

Discrimination indices with negative values are unacceptable since it would mean that a greater number of students in the lower group are choosing the correct answer than in the upper group. Effective distracters, however, are required to have a negative discrimination.

In the present study item discrimination was discussed together with its relation to the item difficulty index (or p-value).

ALTE Materials for the guidance of test item writers suggest (2005:67) that the evidence gathered by CCT 'can only be interpreted in relation to the candidates who took that particular test' and suggest the use of Rasch analysis for further item analysis. Evidence obtained by means of Rasch analysis creates a ground for item calibration, anchoring and creating an item bank. Thus, in the present research Rasch analysis was applied in order to compensate for the limitations of Classical Test Theory.

Rasch analysis is referred to as "a one-parameter item response model", which "uses the data from candidate responses to test items to point a single dimension of measurement, typically language ability (Lumley and Brown, 2005:839).' In comparison to CTT, Item Response Theory (IRT) concentrates not only on the score obtained in the test, but on the concept to the measured. IRT relates "the value of the latent variable to the probability of a correct response (Council of Europe, 2003:107)", thus showing the relation between a test taker's ability and the probability of his/her providing a correct response to an item. In his discussion of IRT Bachman points out the following advantages of the theory (2004:142):

- Item parameter estimates are independent of the group of examinees used;
- Test takes ability estimates are independent of the particular set ot thest items used;
- Precision of ability estimates are known.

Rasch analysis was used was define the amount of the latent trait each item measures and whether the difficulty level of the items corresponds to the test-takers' ability level. Person-item maps were used for this purpose. Such maps present distribution of the persons and items along the variable (the latent trait being measured). As Rasch model uses the same interval measure to evaluate test-takers' ability and item difficulty, they can be placed along the same scale. The measurement unit used in the model is called 'logit'. The origin of the term 'logit' is connected to the 'logarithmic process that is used to estimate item difficulty (ALTE, 2005:69). An item of average difficulty has the measure of zero, items with a positive sign have above-average difficulty while items of below-average difficulty have a negative sign (McNamara, 1996:165).

On the peron-item maps items and persons are ranged according to their difficulty and ability respectively. Thus, the most difficult items and the most able students are located at the top of the scale and the easiest and the least able students at its bottom. A person located at the same logit level as an item has a 50% chance of getting the item right. If an item is located above the student's ability the probability of the correct response reduces, and a test-taker has a greater probability of answering an item correctly if it it located below his measure on the scale. According to the ALTE Materials for the Guidance of Test Item Writers (2005:71), 'ideally' the distribution of item difficulties will mirrir the candidate abilities if the test is to be considered appropriate in terms of the degree of difficulty.'

Person-item maps were also used to define whether there were items in the exam parts which were measuring the same amount of the latent trait and, thus, were superfluous. On the person-

item map these would be the items located at the same logit measure level. Gaps in measurement were also taken into account as they can influence the measurement precision.

Attention was also paid to the difficulty order of the items, which is important in the process of item calibration and standards setting. According to Wright and Stone (1999), when constructing items, a test user makes use of his/her understanding of the variable to be tested, thus, predicting their difficulty. This requires clear understanding of the construct being measured and the way it operationalizes. Therefore, "the difficulty order of items defines the variable's meaning and hence it content and construct validity (ibid.:171).'

The second type of validity evidence inferred from item response data was acquired through the investigation of item fit, which is a useful tool for establishing construct validity. The purpose of fit analysis was to determine the extent to which the observed scores met the scores expected by the Rasch model.

Unidimensionality is the core feature of Item Response Theory and an assumption is made that all items in the test are measuring the same variable. Thus, applying Rasch analysis and investigating fit statistics we can conclude whether all items in, for example, reading part measure the same trait or additional knowledge or skill are required. Rasch analysis provides data on how each item response corresponds to the created model and helps define whether all items are measuring the same trait.

Two types of fit statistics are taken into account when performing analysis: infit and outfit.
Infit refers to "inlier-sensitive or information weighted fit" (Linacre, 2002:878). This means that infit statistics gives weight to the performance of those people who are closer to the item measure (Bond and Fox, 2007:57). If a student fails to give the correct response to the item whose difficulty measure is close to his/her ability measure, it could point to the flaws in the item quality.

At the same time outfit statistics is "outlier-sensitive" (Linacre, 2002:172). Outfit statistics is sensitive to "outlying, off-target, unexpected responses (Bond and Fox, 2007:53)." In a nutshell, outfit statistics is used to define whether a weaker students gave the correct response to a difficult item (guessing, cheating) or a stronger student failed to answer an easier item (carelessness, insufficient time, etc.) Linacre (2002), McNamara (1996), Bond and Fox (2007) state that infit values are more important for evaluating item misfit than outfit values. Thus, McNamara (1996:180) mentions that "outfit problems are less of a threat to measurement than infit ones", while "the infit statistics are the ones usually considered the most informative (ibid.:172)." In other words, infit statistics is connected with the quality of the item as such while outfit statistics would point to the test-taker's unexpected responses and are less threatening to measurement.

Mean square statistics was used to evaluate item fit in the present research. Items showing good fit are supposed to have the value of 1 for mean square statistics. However, Wright and Linacre (1994:370) state that "though the ideal for measurement construction is that data fit the Rasch model, all empirical data depart from it to some extent." The extent to which the data misfits the model is essential for data interpretation. Thus, Wright and Linacre (ibid.) report the following ranges for mean-square fit interpretation, which were also used in the present study:

- >2.0 – distorts or degrades the measurement system;
- 1.5 – 2.0 – unproductive for construction of measurement, but not degrading;
- 0.5 – 1.5 – productive for measurement;

- <0.5 – less productive for measurement, but not degrading; may produce misleadingly good reliabilities and separations;

Items having the value of above the mean indicate unpredictability (it is the item underfits the model), while items having the value below the mean are too predictable and overfit the model (Linacre, 2002:878). Following recommendations provided by Linacre (2010:493), high mean squres were analysed before low ones as they pose greater threat for measurement. High infit mean square were analysed before high outfit; possible reasons for misfit were stated. It should be noted, however, that Rasch analysis presents a mathematical model which does not guarantee construct validity. Characteristics of every item should be analysed and taken care of.

Reading

The initial analysis of psychometric characteristics is provided in the table below as suggested by Kaftandjieva (2010:43).

Parameters		Sample
Number of examinees		22638
Number of items		30
Difficulty	Minimum	0%
	Mean	45%
	Maximum	100%
Discrimination index	Minimum	0.09
	Mean	0.56
	Maximum	0.81
Test score	Maximum	30
	Mean	13.38
	Standard deviation (SD)	6.71
Reliability		0.88
Standard error (SEM)		2.34

Table 1 Psychometric characteristics of Year 12 Exam in English 2010 Reading part

The data show that the test part has a wide range with the largest observation of 100% and the smallest observation of 0%. Thus, the exam part is representative of a wide range of abilities. The mean for the part is quite low about 45% which points to its considerable difficulty. The histogram below points to the same fact:

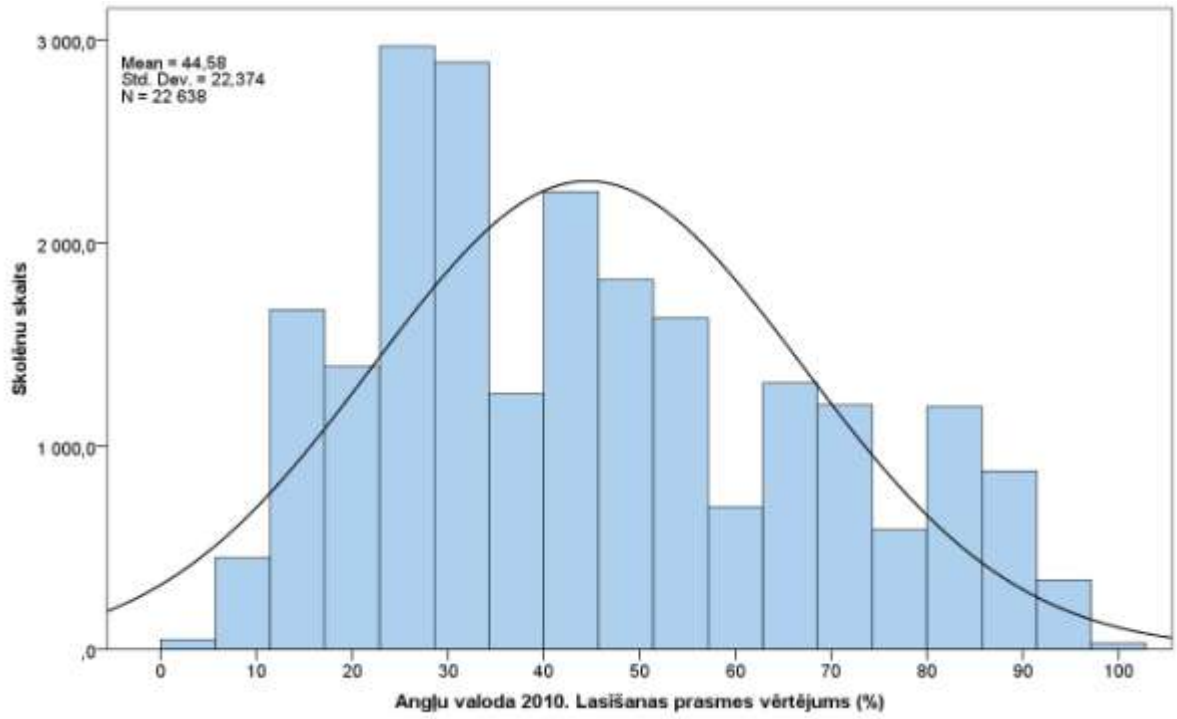


Figure 1 Year 12 English 2010. Reading part score distribution (available at www.visc.gov.lv)

As we can see, the distribution is positively skewed, it is, the majority of scores are in the lower achievement region. The mode of the part is around 25% as most students managed to do only 25%-30% of the part. Nevertheless, the results have a good spread with the standard deviation of 6.71 points or 23%. The part also has a high reliability index of 0.88. However, it is the discriminating power of the items which raises most concerns, especially the lowest value of 0.09. Thus, additional analysis of test items is required in order to single out those items which contaminate the overall result.

	101	102	103	104	105	106	107	108	109	110
Difficulty index	0.57	0.2	0.54	0.44	0.16	0.37	0.43	0.53	0.57	0.54
Discrimination index	0.65	0.18	0.76	0.79	0.20	0.67	0.73	0.76	0.67	0.76

Table 2 Year 12 English 2010. Reading Task 1

The table shows that eight items have acceptable difficulty level ranging between 0.37 – 0.57. Two items (2 and 5) are too difficult for the given population with the difficulty indices 0.2 and 0.16; these are also the items with the lowest discrimination, which makes them less useful for measurement. The rest of the items demonstrate very good discrimination and are able to differentiate between students of different abilities.

	201	202	203	204	205	206	207	208	209	210
Difficulty index	0.48	0.44	0.4	0.82	0.34	0.36	0.75	0.58	0.32	0.63
Discr. index	0.48	0.44	0.29	0.35	0.32	0.35	0.53	0.20	0.09	0.38

Table 3 Year 12 English 2010. Reading Task 2

The difficulty level of all items in Task 2 is within the acceptable range with no items being too difficult or two easy; however, the discrimination indices are considerably lower than those in Task 1. This can be explained by the format of the task (true/false/not mentioned) which allows guessing and, thus, can distort the measurement. Items 201, 202 and 207 provide good discrimination, items 204, 205, 206 and 210 have reasonably good discrimination, items 3 and 8 need to be improved while item 209 has an unacceptable value of 0.09. Thus, even though nine items are capable of differentiating among the stronger and the weaker students, overall discriminating ability is lower than that in Task 2.

	301	302	303	304	305	306	307	308	309	310
Difficulty index	0.52	0.32	0.28	0.4	0.34	0.41	0.47	0.4	0.32	0.42
Discrimination index	0.81	0.6	0.64	0.63	0.67	0.64	0.73	0.72	0.69	0.7

Table 4 Year 12 English 2010. Reading Task 3

According to the data, all items in the task show very good discrimination and are useful for measurement. The difficulty level is also within the acceptable range, though some items are rather difficult (302, 303, 305, 309). However, their difficulty level does not affect their discriminating ability. In general, task 3 includes the best working items in the part.

This fact is also supported by the analysis of Cronbach's Alpha for each task. As the below table shows, Task 3 is the most reliable task in the whole reading part:

Task	Cronbach's Alpha
1.	0.79
2.	0.45
3.	0.84

Table 5 *Reliability of reading tasks*

Additional information about the psychometric characteristics of the reading part was collected by means of Rasch analysis. Firstly, item distribution map was used to examine the item difficulty order. The task is relatively difficult – average item difficulty measure (M) is higher than the average person ability measure, the grouping of items and person abilities also points to the difficulty of the task: person ability distribution is positively skewed.

Most items are grouped around the mean and within one standard deviation from the mean (S). There is a greater variability in the person ability distribution than in the item difficulty distribution. A considerable amount of students, whose measure is below -1, can respond to only three items in the part, the remaining 27 items are obviously too difficult for them. Gaps between items 204, 207 and 210 point to the fact that

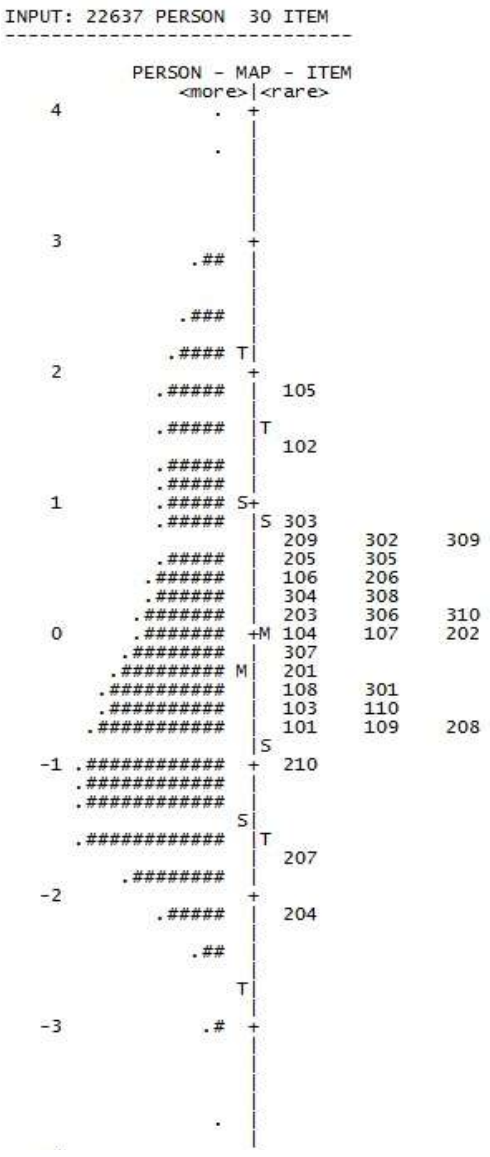


Figure 2 Person-item map for the reading part

a considerable increase of ability is required to move from one measure to another; the same is true for the higher ability students: items 102 and 105 are the most difficult items in the test and there are considerable gaps between them and the rest of the test. There is also a group of higher ability students above the measure of 2 who have no items targeted on their ability.
As far as the targeting in general is concerned, items within one standard deviation from the mean are well targeted and give useful insight into the students' level of reading skills. Items 102, 209, 302, 309 and 207 are less targeted, which could affect ability estimate precision. There are also items which are measuring the same amount of skill, thus, the part could be either shortened by 13 items, or the items could be revised to fill in the gaps at the lower and the higher levels. The distribution also shows that a task can include items of different levels.While Task 3 is consistently measuring above-the-average skills, Tasks 1 and 2 include a wider range of abilities.

Analysis of fit statistics (Appendix 1) allows to judge whether all items in the part are measuring the same variable. The data show that none of the items have infit mnsq value above 1.5, which can degrade measurement. Nevertheless, five items (209, 102, 105, 208, 203, 205) include additional noise as they have infit mnsq values between 1.32-1.46.

Analysis of outfit mnsq (which shows how predictable test-takers' behaviour was) shows that there was considerable randomness in the students' answers. However, the most misfitting items are in Task 2, which can be explained by the format of the task (True/False/Not mentioned). Additional information about the quality of items and test-takers' behaviour can be collected by examining item characteristic curves (ICC). ICC presents a relation between the probability of correct response and the test-taker's ability. The probability of correct response should increase with the increase of ability. If the continuity of ICC is broken, it can help define those areas where the randomness is most pronounced.

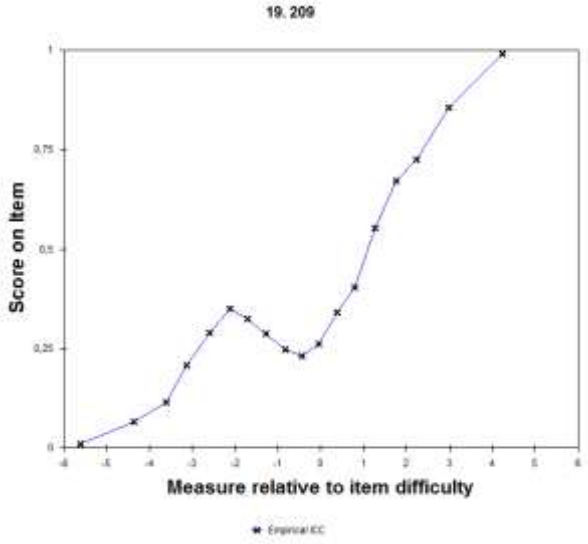


Figure 3 Item characteristic curve for Item 209

Analysis of ICC for the most misfitting item of the part (item 209) shows that unpredictable behaviour occurred in the lower-ability area as the probability of correct response decreased with the increase in ability. Such items require more thorough analysis and examination.

The following conclusions have been drawn from the analysis of the reading part:

- test items measure a wide range of abilities even though overall difficulty of the part exceeds the level of the target population's skills; however, there is still a number of higher-ability students whose skill level has not been targeted;
- in general, the test part has a good overall discriminating power, though there are items with unacceptably low discrimination indices. These are mostly found in Task 2, whose format gives room to guessing;
- even though reliability index for Task 2 is very low, the whole part is quite reliable with Cronbach's alpha of 0.88;
- Rasch analysis shows that the test is mostly targeted on the students of average and above-average ability as most items are located in this area; however, the test does not quite efficiently measure highest-ability students. Most items measure the same amount of skill;
- Fit statistics show that items mostly measure the same variable. Nevertheless, it should be noted, that the most misfitting items are those which go beyond lexical knowledge and require making inferences and drawing conclusions about the text read.

Language Use

Parameters		Sample
Number of examinees		22637
Number of items		45
Difficulty	Minimum	0%
	Mean	46.75%
	Maximum	100%
Discrimination index	Minimum	0.17
	Mean	0.44
	Maximum	0.81
Test score	Maximum	45
	Mean	21
	Standard deviation (SD)	8.36
Reliability		0.89
Standard error (SEM)		2.82

Table 6 Psychometric characteristics of Year 12 English 2010 Language use part

The above table shows that the test part has a wide range with the items covering all ability levels. The mean for the part ir 46.75%, which shows that the part was quite difficult.

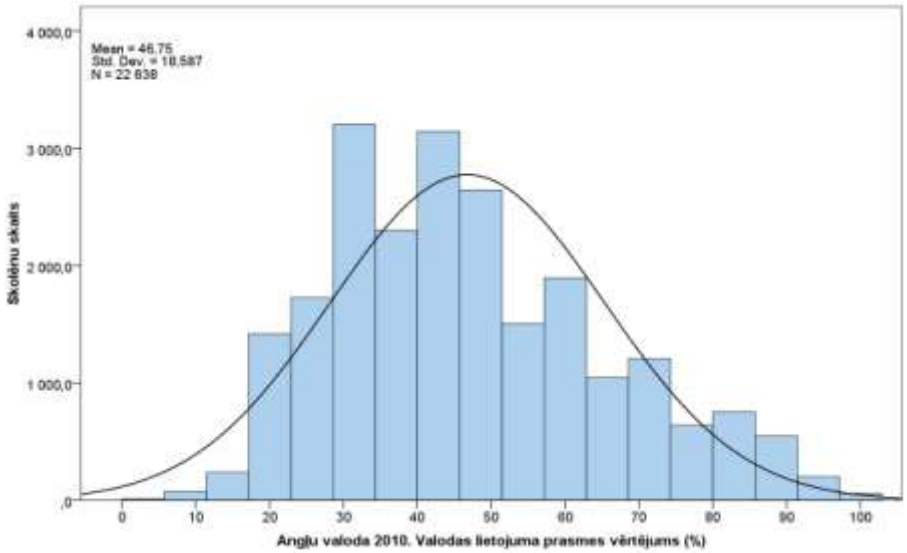


Figure 4 Year 12 Exam in English 2010. Language use part score distribution

The standard deviation of 18,587% also shows that the scores are spread out over a sufficient range of values. However, the mode (around 30%) and the median (around 44%) point to the relative difficulty of the part. The distribution is positively skewed and the bulk of scores lie to the left of the mean in the lower score region. It should be noted that positively skewed distribution has a lower discriminating ability for the students whose scores are below the mean. Thus, we can assume that the language use part is useful for separating the weaker students from the stronger ones, but it is not informative enough when differentiating among the students in the weaker group.

The overall reliability index of 0.886 is reported for the part, which makes the measurement quite reliable. If analysed indivudually, the tasks show the following reliability indices:

Task	Cronbach's alpha
1	0.74
2	0.49
3	0.85

Table 7 Reliability of the language use tasks

Task 3 (open cloze) is the most reliable in the part while Task 2 (error correction) is the most unreliable. However, low reliability index of Task 2 can be largely connected to the scoring procedure rather than to the quality of the task itself. Marking the lines of the task as correct (√) or incorrect (-) if they contained an extra word instead of providing the word in question gave the test-takers fifty per cent chance of guessing the answer. Had the students been asked to provide the eliminated word instead of just marking the line as incorrect, the scores for the task would have changed drastically. Thus, during the analysis account should be taken not only of the task quality but also of the quality of its scoring techniques as it can affect task reliability.
Analysis of individual items has given the following results:

Item	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118
p-value	.66	.59	.25	.46	.58	.55	.47	.25	.26	.35	.47	.37	.50	.39	.55	.74	.52	.27
DI	.28	.44	.32	.41	.41	.75	.58	.46	.22	.45	.56	.43	.63	.28	.44	.28	.56	.18

Table 8 Language Use. Task 1

The mean for task 1 is 45.77, four items (3, 8, 9, 18) prove to be too difficult for the target group. None of the items are too easy. The four most difficult items refer to different areas of knowledge: item 3 – the use of 'hardly any', item 8 – collocation 'bear in mind', item 9 – subordinate clause in the second conditional, item 18 – the use of bare infinitive in complex object.

12 items out of 18 show acceptable discrimination index value, six items (1, 3, 9, 14, 16, 18) are not efficient enough for separating weaker students from the stronger ones. The best discriminating items are 6 (0.75) and 13 (0.63). Analysis of these items shows that they require the use of the language in context rather than formal knowledge of grammar rules.

Low discrimination of items 3, 9, 14 and 18 can be attributed to the difficulty of the items. However, closer analysis of alternative statistics and distracters shows that each of the above mentioned items has two answer options with the positive point biserial correlation. Thus, in these items even stronger students were opting for the wrong answer.

Item 103 – negation / double negation

He was a big man with _____ neck, although he did have a very large moustache.

A hardly any B hardly some *C hardly no* D hardly such

Key *next option*

Item 109 – Second conditional

They didn't think they could bear it if anyone _____ about the Potters.

A would find out B find out **C found out** D had found out

next option **Key**

Item 114 – Second conditional

The Dursleys shuddered to think what the neighbours would say if the Potters ____ in the street.

A arrived B arrive *C would arrive* will arrive

Key *next option*

Item 118

None of the Dursleys noticed a large tawny owl ____ past the window.

A to flutter **B flutter** *C fluttered* D would flutter

Key *next option*

Item 14 shares the characteristics of item 9 as it is based on the same grammar rule (the second conditional). In general, the discriminating ability of the task is sufficient but not high enough, as most items have the index value slightly above 0.40.

Item	201	202	203	204	205	206	207	208	209	210
p-value	0.74	0.83	0.65	0.87	0.47	0.47	0.69	0.76	0.75	0.79
DI	0.43	0.29	0.34	0.26	0.35	0.17	0.41	0.35	0.18	0.36

Table 9 Language Use. Task 2

According to the above provided data the task was quite easy with the mean score of 70.2%. However, the results of the task should be treated with caution as the reliability index for the task

is only 0.49. The present data show that four items (2, 4, 6 and 9) have unacceptably low discrimination indices, only two items (1 and 7) have sufficient discriminating ability.

Item	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317
p-value	.18	.66	.19	.21	.69	.37	.19	.32	.38	.16	.22	.25	.14	.41	.34	.53	.58
DI	.40	.65	.37	.49	.57	.65	.44	.51	.74	.33	.36	.59	.39	.62	.65	.81	.47

Table 10 Language Use. Task 3

Task 3 is the most difficult task in the part with the mean score of 33.05%; however, it is also the best discriminating task as none of the items has an unacceptably low discrimination index value. Even the most difficult items (1, 3, 7, 10 and 13) are capable of discriminating between the stronger and the weaker students.

Rasch analysis of the task allowed to draw the following conclusions:

- Person-item map (see Figure 3.2.2) shows that the test-takers' ability is slightly below the difficulty level of the part; however, in general, the two are quite mirrorred;
- Both item difficulty and persona abilities show a good spread, test items account for almost the whole range of abilities leaving out a small group of the strongest and the weakest students.
- There are no items which would be too difficult or too easy for most test takers. However, there are gaps in the distribution of items below the mean. This means that lower ability students need a considerably bigger increase of skill in order to move from one score to another.

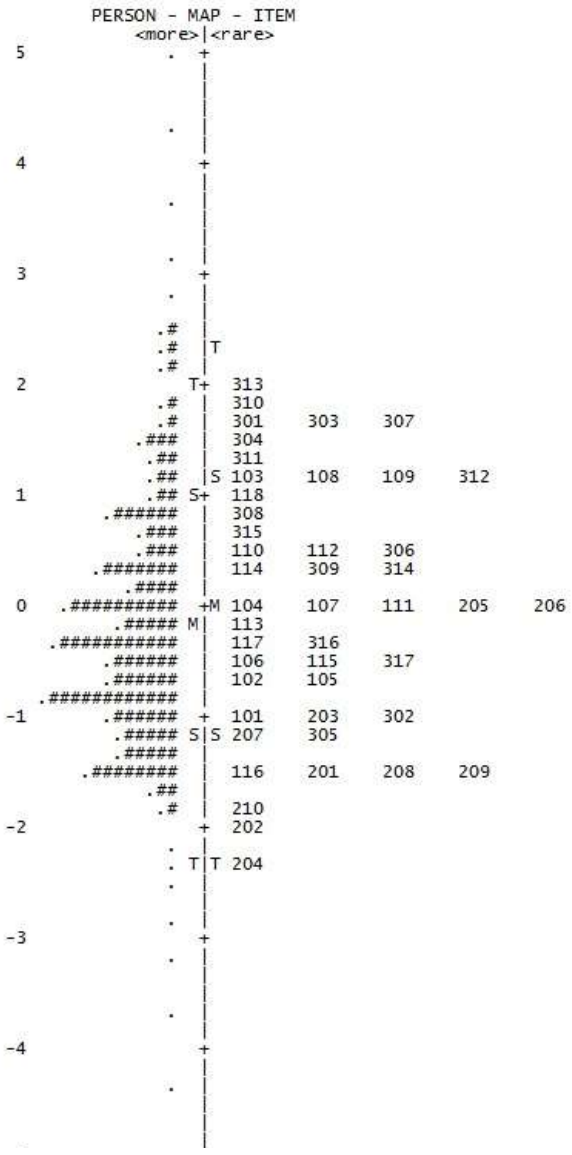


Figure 5 Person-item map for the language use part

As far as the item targeting is concerned, items 313 and 202 are the worst targeted items since they do not correspond to a particular group of students. Item 204 is targeted on a small number of lower ability students, at the same time it is too easy for the majority of students and too difficult for the rest of the low ability group.

The majority of items are measuring the same amount of skill, thus, the exam part could be shortened or the items could be revised to fill in the gaps in measurement.

Task 3 proves to have the most difficult items with the majority of them assessing higher ability students within 2 st.d. from the mean. Task 2 proves to have the easiest items, however, this fact should be treated with caution taking into account low reliability of the task.

Analysis of fit statistics (see Appendix 2) shows that none of the items have unacceptably high infit mnsq values, thus, the principle of unidimensionality is largely observed. Item 206

contains the biggest amount of noise (30%). Four items (209, 118, 206, 109) point to randomness in the students' behaviour. It should be noted, though, that the most misfitting items are found in tasks 1 and 2, whose format allows guessing.

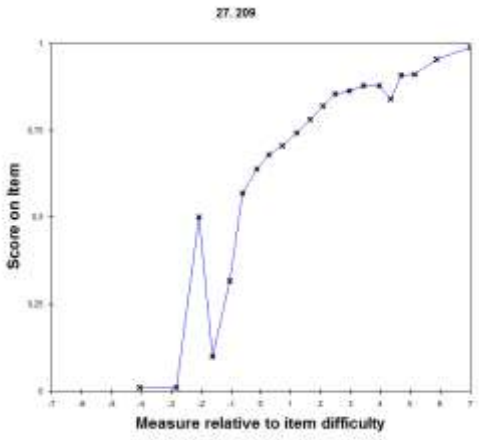


Figure 6 ICC for item 209

Item 209 is easy (-1.54) with an acceptable infit mnsq value (1.15) and a high outfit value (1.62).

High misfit is obviously connected with stronger students failing on an easy item. However, it is impossible to provide item option analysis since we have no information about the choice the stronger students made.

Lower ability students have a high probability of getting a correct response (the line contains no mistakes). However, this choice can be attributed not only to their actual knowledge but to guessing/cheating or also inability to find mistakes in the text in general and ticking many lines as correct. Students within the band of -2 to -1.5 have a lower possibility of providing a correct response (probably finding a 'mistake' in the correct line or making a guess and marking it as incorrect). There is also a surprising drop in correct response probability in the higher ability region which is difficult to explain since we do not know what words in the line were crossed out.

All in all, we can conclude that despite a certain amount of noise in items 118 and 206, the items are measuring the same trait. Test-takers are mostly demonstrating random behaviour in tasks 1 and 2. In Task 1 items with the highest outfit value are based on the knowledge of definite grammar rules (second conditional, bare infinitive in complex object).

The following conclusions can be drawn about the quality of the language use part:

the difficulty level of the task is above the target population's ability level: the distribution of test scores is positively skewed. 50 per cent of the scores lie below the median of 44%;

the overall discriminating ability of the examination part is acceptable (0.44) but not high enough, the best discriminating items are in task 3;

the results of the examination part are reliable (0.89), the least reliable task is task 2, although its reliability has been affected by the scoring procedure;

task 3 is mostly measuring the students whose language proficiency is above the average, task 1 covers a wider range of ability. Task 2 is the easiest task as most of its items are located in the lower skill region. However, this result should be treated with caution due to the low reliability index of the task;

most ability levels are well targeted apart from a small number of the highest ability students;

fit statistics shows that on the whole items are measuring the same variable; nevertheless, items in task 1 and 2 have high outfit measures which are mostly produced by lower ability students scoring on the items above their measure.

Listening

Parameters		Sample
Number of examinees		22641
Number of items		30
Difficulty	Minimum	0%
	Mean	55%
	Maximum	100%
Discrimination index	Minimum	0.08
	Mean	0.47
	Maximum	0.83
Test score	Maximum	30
	Mean	16.53
	Standard deviation (SD)	5.73
Reliability		0.84
Standard error (SEM)		2.30

Table 11 Psychometric characteristics of the listening part 2010

The listening part is the easiest in comparison to the reading and language use parts. The whole range of scores ir presented (from 0% to 100%). The histogram below shows that the distribution is quite symmetrical with the majority of scores lying in the middle part. The scores are well spread, though there is a noticeable cluster in the area between 40% and 58%.

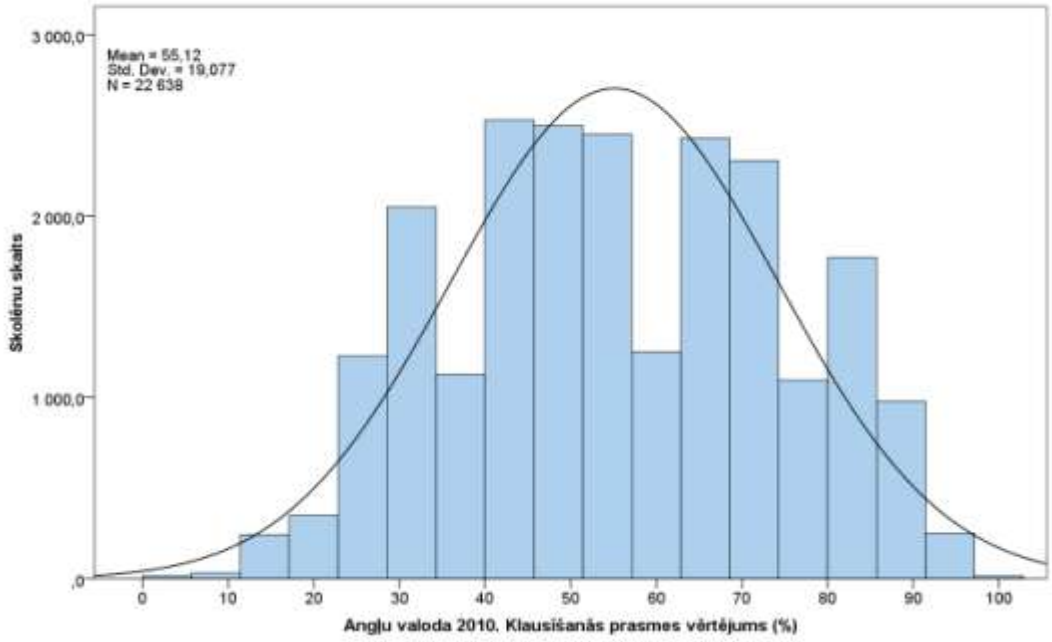


Figure 7 Year 12 Exam in English 2010. Listening part score distribution

The reliability index for the part is quite high – around 0.84. However, not all tasks of the part are equally reliable:

Task	Reliability
1	0.84
2	0.50
3	0.49

Table 12 Reliability of the listening part 2010

The above table shows that when examined separately, only one task has sufficient reliability index. This could be attributed to the format of tasks two and three (True/False and multiple choice, respectively).

The mean discrimination index for the exam part is noticeably low (0.47) with the lowest index value of 0.08. Tables below allow to examine each difficulty level and discrimination index of the part separately:

Item	101	102	103	104	105	106	107	108	109	110
p-value	0.53	0.35	0.43	0.85	0.76	0.79	0.52	0.54	0.40	0.45
DI	0.80	0.63	0.71	0.41	0.55	0.58	0.83	0.69	0.72	0.80

Table 13 Listening 2010. Task 1

The above table shows that all items in task 1 have good discriminating ability and their DI values are higher than 0.40. The format of the task excluded guessing, which definitely influenced its results.

Item	201	202	203	204	205	206	207	208	209	210
p-value	0.95	0.59	0.15	0.83	0.56	0.73	0.39	0.66	0.77	0.66
DI	0.08	0.44	0.21	0.36	0.32	0.43	0.51	0.53	0.27	0.40

Table 14 Listening 2010. Task 2

The quality of item 203 should be looked into as it is quite difficult and has an unacceptably low discrimination index. Closer investigation of the item shows that the difficulty of the item was determined by the similarity in pronunciation between ‘Arctic’ and ‘Antarctic’, which were the key to the item:

Item 203
The scientist had been working in the **Arctic** for 6 months. – False
Tapescrips: Apparently, Dr Patso had been working in the **Antarctic** for 6 months when she came across a new species.

Low DI of item 201 can be attributed to its easiness (p-value of 0.95). The resto of the items have acceptable difficulty level; however, discrimination indices are acceptable but quite low.

Item	301	302	303	304	305	306	307	308	309	310
p-value	0.61	0.23	0.58	0.22	0.58	0.51	0.73	0.51	0.23	0.41
DI	0.72	0.22	0.41	0.26	0.31	0.53	0.30	0.39	0.32	0.39

Table 15 Listening 2010. Task 3

Only three items of the task have good discriminating ability (301, 303 and 306). Items 302 and 304 are not effective for measurement. The rest of the items have acceptable but not high enough DI values.

Thus, Task 1 is the best discriminating task in the part as its discrimination indices range between 0.41 – 0.80. Task 2 contains the worst discriminating item of the part with the index value of 0.08.

Rasch analysis shows that generally the difficulty level of the listening part corresponds to the students' ability level as their mean measure is slightly higher than the item difficulty measure.

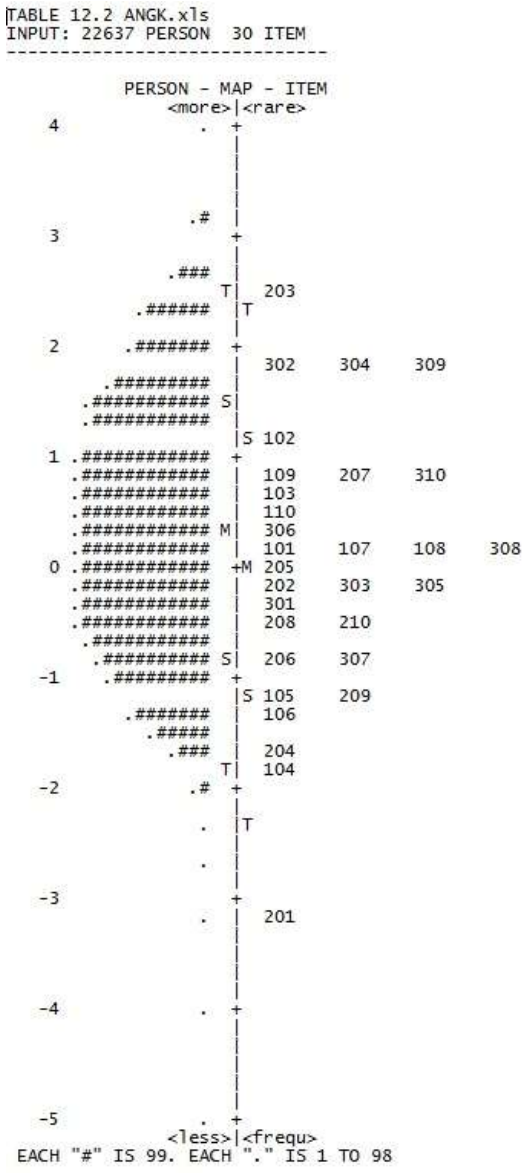


Figure 8 Person-item map for the listening part

Both person measure and item difficulty have symmetrical distribution and task items have similar spread (person standard deviation (SD) of 1.01 and item SD of 1.19). Thus, variability in the students' ability corresponds to the variability in item difficulty. On the whole, item distribution mirrors person distribution.

17 items are clustered in the band within 0.9 – (-0.7) logits, measuring average ability students. Items in this area are well-targeted; however, there are items which are testing the same amount of skill and could have been revised to fill in the gaps in measurement in other areas, for example in the higher and the lower skill region.

Higher ability range is assessed by only 4 items (203, 302, 304, 309) which, however, do not provide precise measurement. More varied items could be introduced to differentiate among the more able students. Gaps in measure between items 102, 203 and the cluster of 302, 304 and 309 show that a considerable increase in skill is necessary to move from one measure to another. Item 201 is the easiest in the exam part and assesses a limited number of people.

Task 2 assesses the widest range of abilities, items in task 1 also have a good spread though they do not account for the lowest and the highest abilities. Items of task 3 are targeted on the students of average and high ability.

Analysis of fit statistics (Appendix) shows that all items of the listening part have infit mnsq within the acceptable range, which proves that all items all testing the same skill (are unidimensional) and are useful for measurement. However, there are items that show high outfit mnsq and, thus, point to unexpected behaviour. The three most misfitting items are also the most difficult items in the test (302, 203, 304).

Additional analysis of item characteristic curves shows that in item 302 unexpected behaviour happens in the lower ability area (measures -4 - -1). The probability of the correct response fluctuates and the continuity of its increase is broken.

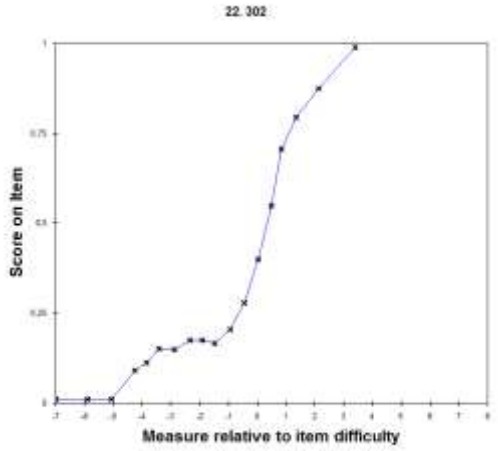
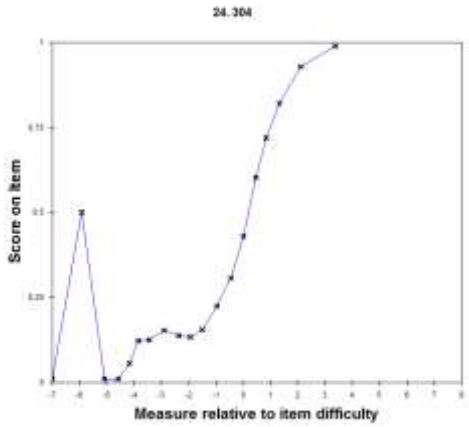


Figure 9 ICC for item 302

Analysis of the tapescript shows that these students could have been mislead by the wording of the item and the tapescript. However, this fact does not diminish the quality of the item as more able students whose measure was above the average were able to choose the right option (probability of the correct response is growing in the range between -1 and 4):

Item 302 The purpose of the special arts programme in Rhode Island was to: C investigate the influence of arts training. - Key D study the <u>impact</u> of music and arts on math.

Tapescrips: 'We started out wanting to see the <u>impact</u> of arts training in some first and second grade kids.'



High outfit for item 304 can be explained by the students' unexpected behaviour at the lower region of skill, which is quite pronounced in the ICC.

Figure 10 ICC for item 304

In general, this is a difficult item with high discrimination above the measure of -2 (the slope of the curve is quite steep). However, students with the lowest ability (measure -7 – (-6)) also had a high probability of choosing the correct response. This could be attributed either to cheating or guessing linked to the wording of the item and the tapescript.

Item 304

During typical lessons students:

A went to concerts and talked about music.

C mostly *listened* to music. – Key

Tapescript: 'The typical music lesson tended to be somewhat passive', says Gardner. Students *listened* to tapes and concerts and talked about music in class.

Thus, items showing highest outfit measures are in tasks 2 and 3. It can be concluded that these are the items which require not only 'hearing' the right information but also deeper understanding of its meaning. Students were easily mislead by similar wording of the answer variants.

The following conclusions can be drawn about the listening part:

1. The diffiuculty of the examination part corresponds to the test-takers' abilities.
2. Overall discriminating ability of the part is acceptable (0.47), even though there are items (especially in task 2 and 3) which are not effective for measurement.
3. Pearson-item map analysis shows that in general items mirror test-takers' ability measure, however, most of the items assess students in the average ability area. A number of items which measure the same amount of the latent trait could have been reworked to fill in the gaps in the higher ability region.
4. Fit analysis shows that all items are measuring more or less the same trait as the amount of noise does not exceed 20%. However, there is a number of items having high outfit measures, thus, pointing to the students unexpected behaviour. Analysis of ICC for the most misfitting items shows that randomness appears in the lower skill region with less able students providing correct responses to difficult items.

Correlation analysis

Analysis of individual examination part scores was complemented by the analysis of the correlation between the examination parts as it was essential to establish their relationship and consistency of measurement. The analysis helped define whether all exam parts provide the same distribution of levels, which could be linked to the CEFR. If there is high positive correlation between the parts, then we can assume that these parts measure 'different aspects of the same construct of language use, and this reflect a common underlying ability (Bachman, 2004:108).' We can also assume, that if the link between the CEFR and a part of the examination has been established, the correlation of the part and the rest of the tasks can point to their correspondence to the CEFR.

Table 16 gives the insight into the relationship between the examination scores in different parts.

English

	Listening			
Reading	**0,787**	Reading		
Speaking	**0,687**	**0,654**	Speaking	
Language use	**0,769**	**0,802**	**0,666**	Language use
Writing	**0,706**	**0,688**	**0,757**	**0,727**

Table 16 Correlation between the examination parts (Year 12, 2010) available at www.visc.gov.lv

According to the above data the correlation coefficient ranges from 0.654 to 0.802, which points to a considerable amount of consistency of the examination scores. We can conclude that there is a strong relationship between language skills of the same type: reading and listening (0.787) as receptive skills and writing and speaking (0.757) as productive skills. The rest of the parts are also demonstrating sufficient strength of relationship. Thus, taking into account the fact that the writing and speaking parts of the examination have the strongest link to the CEFR (see the present research), we can assume that the remaining parts also follow similar principles of level distribution.

Conclusions

The present research allows to conclude that the three analysed parts (listening, reading and language use) produced scores effective for assessing the students' language proficiency. The examination parts are reliable, even though individual tasks produce unacceptable reliability indices when analysed separately. The most unreliable are those tasks which allow a certain amount of guessing.

The difficulty level of the listening part corresponds to the ability level of the test-takers. Reading and language use parts are obviously quite difficult for the target population, which is proved by both descriptive statistics and Rasch analysis. However, Rasch analysis also shows that the items of all three parts are mostly targeted on the average ability students. Thus, a number of items should be introduced to give more precise measurement of the highest ability students. In all parts there are items which are measuring the same amount of the latent trait and, thus, could have been reworked to fill in the gaps in measurement. At the same time, item distribution shows that the ordering of items according to the difficulty is not connected to the number of the task. For example, in the reading part the most difficult items are found in task 1 and not in task 3 as it would be expected. The easiest items in the said part are found in task 2. Thus, each task of the parts covers a wide range of abilities. At the same time Rasch analysis shows that the examination divides the target population into a number of strata which could be described in terms of the CEFR levels. However, this would require individual item calibration and setting of the cutscores. Thus, when comparing exam tasks to the CEFR account should be taken of each individual item and not of the task in general.

Correlation analysis allows us to establish the relationship between the examination parts and proves that they all are based on the same underlying principle. Strong relationship between speaking and writing, whose marking scales are linked to the CERF levels, gives additional validity to the interpretation of the scores. At the same time, considerable correspondence between speaking and writing and the rest of the parts allows us to assume that they are based on the same principle and, thus, can be linked to the CEFR.

All in all, items in the parts are based on the same trait/skill as the observed results comply with the expected scores. Randomness in the students' behaviour is mostly observed at the lower level with students giving correct responses to the items above their ability.

It can be concluded that Year 12 Examination in English provided useful and reliable assessment of the test-takers' language proficiency level and is a valid measurement tool.

Materials and methods applied in the study demonstrate the potential of IRT in general and Rasch analysis in particular in the process of test development and analysis and prepare the ground for further research in the field. Item difficulty measures obtained in the study can be used in the future development of examination tasks for item anchoring and creation of an item bank.

Qualitative relation procedures

Formal foreign language testing has not been a sensitive issue in Latvia up to the moment when our country has become a member of the European Union. Previously, foreign language examinations, in our case of interest the English language, were a matter of domestic education policy of our country, which determined the levels to which it was possible to know the foreign language and the examination procedures. It was even less complex to develop all the standards due to the fact that Latvia is a multicultural country and, for example, the levels of second language proficiency - Latvian for many - could be transferred to the foreign language proficiency level description.

After Latvia has been granted the membership of the European Union, Latvian legislators and educators found themselves in a difficult situation. Firstly, European Union dictated their laws and standards as concerns foreign language proficiency – new tests, new proficiency levels, and alike (generally known as Common European Framework of Reference for Languages – CEFR). In order to maintain the new status, Latvia had to follow the general regulations applicable for all countries. Our country was interested in presenting itself as a full-fledged member of the European Union; therefore, it was necessary to deal with the issue as soon as possible. Secondly, the standards offered by the European Union displayed certain difference in comparison with Latvian standards. Current examinations were and still are to be related to the CEFR standards. The examination and assessment system were to be changed dramatically according to the European standards, which meant more work for legislators, educators and language learners. Thirdly, Latvia was interested in intelligent, competent professionals who would be appreciated world-wide, which, naturally, was not possible without the sufficient language proficiency level accepted in the European Union. The greatest difficulty in this situation is represented by the still ongoing discussion about the issues such as the structure of the English language examination, the degree to which the CEFR language proficiency levels are to be merged with the existing Latvian standards, and alike. Currently Latvian state examination of the English language in the 12th Grade consists of five parts: speaking, reading, writing listening and language use.

The project team for analyzing the Year 12 Examination in English 2010 consisted of twelve Master students of the University of Latvia. They were Natalja Skvorcova, Marina Brunere, Irina Smirnova, Olga Smirnova, Ineta Egliena, Santa Strēle-Ivbule, Alma Bernharda, Zilgme Eglīte, Anna Lavrecka, Aira Misa, Tatjana Savenkova and Margarita Šendo. The work of the team was conducted by Doctor of Philosophy in Linguistics Vita Kalnbērziņa and supported by State Education Centre official Gundega Muceniece.

Our project group used the Manual for analyzing the exam and relating it to the CERF since CEFR is ‘a descriptive instrument’, so is the Manual, as its authors say, ‘depending on how far the linking is intended to be implemented and in what kind of assessment context and culture this is to be attempted, projects will differ in terms of how far they have progressed in verifying of their tests and examinations to the CEFR’” (Figueras et al 2005 in Kalnbērziņa, 2006).

According to the stages of the ‘process of building an argument based on a theoretical rationale’ provided by the Manual, the procedure of describing the relation of Year 12 English language examination to the Common European Framework of Reference included the following stages:

1) the internal familiarization with CEFR;
2) the specification of the examination contents;
3) the standardization of researchers’ judgements based on standardized samples;

4) the empirical validation.

The first procedure to obtain the relation of the examination to the CEFR was familiarisation, which is a selection of training activities designed to ensure that the participants in the linking process have a detailed knowledge of the CEFR, its levels and illustrative descriptors. It is important to make a distinction between familiarization with the CEFR itself and with the rating instruments to be used. During the familiarisation phase all the members of the team discussed and shared their experiences in using the CEFR. Moreover, the team expressed their opinions about Latvian Year 12 examination and its levels.

During the first stage, the internal familiarization, the project group members received the copies of the Year 12 examination in the English language 2010 and analysed them. Besides, the levels in the CEFR were discussed and our experience in their use was shared.

Next, the team moved to the specification procedure. According to Figueras (et al 2009) the specification procedures involve analysing the content of the examination in question, in relation to the relevant categories of the CEFR The specification stage is aimed at producing 'a report on how well the examination content reflects the descriptive categories of the CEFR' (Figueras et al 2005:268 in Kalnbērziņa, 2006). During this stage, the group focused on doing the tasks and finding out the subskills and competences which were tested in the examination tasks.

During the next, standardization stage, we were analyzing the tasks that required both receptive and productive skills. We analyzed and assessed the tasks which have already been assessed by other examiners, comparing and sharing our opinions. The project group also defined the difficulty level of the tasks, as well as their relation to the CEFR.

The last stage of the process was the empirical validation. 'The aim of validation is the collection and analyses of test data and ratings from assessments to provide evidence that both the examination and the linking to CEFR are sound' (Figueras et al in Kalnbērziņa, 2006).

There are six properties as the bases for quality control in test development: *reliability, construct validity, authenticity, interactiveness, impact and practicality*. Reliability is concerned with how accurately the test measures. Construct validity is the extent to which the test measures the right construct. Authenticity is the extent to which the test takers are perceived to share the characteristics of the target-language use tasks. Finally, interactiveness - the extent to which the test tasks engage the same abilities as the target-language us tasks. (Bachman and Palmer, 2001)

A valid test provides consistently accurate measurements, therefore it is reliable. The test developers should understand that "there will always be some tension between reliability and validity. (Hughes, 2003)

	Nr of test-takers	Mean %	St. Dev %
2008	23526	40.28	21.29
2009	23652	42.24	23.56
2010	22638	44.58	22.37

Table 1 ***Comparative Statistics of Year 12 Examination***

As it follows from the statistics shown in Table 1, the situation has not changed significantly over the past 3 years. The mean has grown a little which indicates that the average difficulty level the exam takers can manage has increased by a couple of per cent. By 2010 it has reached 44.58%, which is quite appropriate though it is the lowest compared to other skills – e.g. the mean in Listening is 55.12, in Speaking it is 62.62; the general mean in the whole exam is 51.24 (available at http://visc.gov.lv/eksameni/vispizgl/statistika/2010/stat2010.shtml). This signals that the Reading part of the examination appears to be the most difficult. The tendency over the years has remained the same; the results in Reading are among the lowest.

Here we will present the results of the analysis of each of the five tests: Reading, Listening, Language Use, Writing, and Speaking.

Relation of Year 12 English Language Examination Reading Test to CEFR (Nataļja Skvorcova, Marina Brunere, Irina Smirnova)

The part discusses the relation of Year 12 English language examination to the Common European Framework of Reference (CEFR) for languages: Learning, Teaching, Assessment. The research is based on the analysis of Year 12 exam tasks 2010, reading comprehension, and sample blank Grid and sample specification of a reading test including standardizations of our judgments.

Definition of Reading

The term 'reading' is defined in various ways. Researchers agree that reading is more than mere decoding of orthographic symbols (Nuttall, Tamrackitkun). Nuttall suggests reading is about decoding, or deciphering; it also employs understanding and seeking for meaning (2005: 2). Tamrackitkun investigates various ideas on what reading is, pointing out that it is viewed as gathering or choosing from what was written; the process of gathering information; the active process related to problem solving; requiring both visual and non-visual information and prior knowledge (2010 :14). She stresses reading as interaction between the reader and the text (ibid.: 16).

According to Nuttall, assessment of students' achievement at the end of a course or year is one of the main reasons to test reading (2005: 217). Various tests check the students' literal comprehension, interpretive comprehension and critical reading (Mohamad 1999). Comprehension at literal level 'involves surface meanings' (ibid.). Interpretive comprehension is a higher level of reading comprehension, at this level 'students go beyond what is said and read for deeper meanings' (ibid.). At the next level of critical reading the students are expected 'to differentiate between facts and opinions…, to recognize persuasive statements and to judge the accuracy of the information given in the text' (ibid.). The present paper describes subskills of reading which are being tested in the Year 12 exam 2010.

While testing reading, teachers are interested in checking the following macro-skills:

- Scanning text to locate specific information;
- Skimming text to obtain the gist;
- Identifying stages of an argument;
- Identifying examples presented in support of an argument (Hughes 2003:120).

Hughes mentions the following reading micro-skills:

- Identifying referents of pronouns;
- Using contexts to guess meaning of unfamiliar words;
- Understanding relations between pats of text by recognising indicators in discourse (ibid.: 117).

Further, this study offers information about the subskills which are tested in the reading part of Year 12 exam.

Materials and Methods

The following materials were used for the present study: Each item of the reading comprehension part of Latvian Year 12 English language examination paper for the year 2010 provided by the Ministry of Education was analyzed; *Common European Framework of Reference for Languages* was consulted, the Manual for Relating language examinations to the Common European Framework of Reference for Languages 2003 was used; the statistics of the reading part of exam results of 2008, 2009 and 2010 were examined.

Each task of the test needed analysis according to the guidelines of relating exams to the CEFR and its descriptors discussed in the Paper. Table 2 allows us to get a general impression of the reading part of the Year 12 exam. The reading part of the examination consists of three tasks, requires 50 minutes and is aimed at testing reading comprehension skills.

No of tasks	3
Integration of skills	Reading comprehension
Total test time	50 minutes
Target performance level	B2+
Purpose	General proficiency (comprehension)

Table 16 General Information about the Reading Test

As it can be seen from Table 2, Year 12 examination tasks test general proficiency asking the students to read the texts and complete different tasks in terms of the students' reading comprehension in the 50 minutes provided.

We compared the task demands with CEFR. The first task, 'Annual Sled Dog Race' from English Teaching Forum did not demand high levels C1 or C2, but was focused on the students' ability to recognize the textual schema on the syntactical level. However, task 2, 'The Open Window' by H. Munro and task 3, 'History of Cheese' were more demanding in terms of good knowledge of vocabulary and grammar. Thus, task 2 corresponded to level B2+, while task 3 referred to the higher level C1. Therefore, we included descriptors from the level of C1 to B2 in the scale of overall reading comprehension.

C1	*Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of specialisation, provided he/she can reread difficult sections.*
B2	*Can read with a large degree of independence, adapting styles and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.*

Table 17 Overall Reading Comprehension in terms of the CEFR

Thus, it is clearly seen that task 2 and task 3 are more challenging and demanding, as the students need to have rather high lexical competence and understand the essence of difficult sections for reading comprehension.

Therefore, it is worth examining the difficulty level for reading strategies in different items of each task separately. According to item statistics, all three tasks included some items which demanded more efficient language competence and introduced a number of difficulties.

Task 1 includes ten items which are missing in the text and the students were asked to find the most appropriate part for each item. The most difficult items of task 1 report the following results:

- 5 (16% correct)
- 2 (20 % correct)

These items require such reading strategy as reading for orientation, especially for good orientation in the concept of time and place, which are rather contradictive.

According to the CEFR: learning, teaching, assessment, the level of the task 1 which refers to the level B2 has the following characteristics of reading for orientation:

- B2: *Can scan quickly through long and complex texts, locating relevant details. Can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.*

Task 2 includes ten items, each of which has to be identified by the students as true, false or not mentioned according to the text. The most difficult items of the task 2 have the following characteristics:

- 9 (32% correct)
- 5 (34% correct)
- 6 (36% correct)

These items require the ability to grasp the essential meanings and involve two strategies - reading for information and argument as well as reading for orientation. However, during the analysis of the items in the group of MA students there were discussions concerning this task and several items were regarded as too demanding and challenging because they were not mentioned in the text directly and asked the students to guess the hidden information from the text. These items are as follows: 8, 9 and 10. Therefore, we do not completely agree with the Key 2010 which accepts only one correct answer for each item. Thus, item 9, for example, could have both variants such as false (F) and not mentioned (NM).

Task 3 is considered to be the most difficult referring to level C1. It includes ten items and the students are asked to find the most appropriate part of the sentence to be filled in. Besides, it provides more phrases than needed, therefore, the students have be able to distinguish between the necessary parts and unnecessary and be good at both syntactical and lexical levels.

The most difficult items of the task 3 are as follows:

- 3 (28% correct)
- 9 (32% correct)
- 5 (34% correct)

These items require such reading strategies as reading for orientation and reading for information and argument. The task demands good knowledge of collocations and additional attention to the vocabulary use and syntax.

According to *CEFR: Learning, Teaching, Assessment*, the level of task 1, which refers to level C1, has the following characteristics of reading for information and argument:

- *C 1: Can understand in detail a wide range of lengthy, complex texts likely to be encountered in social, professional or academic life, identifying finer points of detail including attitudes and implied as well as stated opinions.*

Thus, it is assumed that the level of the reading tasks is rather high and the difficulty level of tasks 2 and 3 may have some negative results for those students who do not correspond to level B2, as the tasks are challenging and demand certain knowledge of syntax, grammar and vocabulary.

After examining the level of difficulty in each item and defining the most challenging items of the three tasks in the reading test, it is worth giving characteristics to each text separately according to the CEFR.

Characteristics	**Task 1**	**Task 2**	**Task 3**
Text source	journal	fiction	text book
Authenticity	modified	modified	authentic
Discourse type	descriptive	narrative	descriptive
Domain	public	personal	public/professional
Topic	sport	family, people, relations	food
Nature of content	concrete	concrete/abstract	concrete
Text length	140 words	300 words	260 words
Vocabulary	rather extensive	rather extensive	rather extensive
Grammar	simple/complex	rather complex	somewhat complex
Language level	B2	B2+	C1

Table 18 Analysis of the Characteristics of Reading Test

The table above describes each task in details in terms of the the CEFR: we can see again that the level of difficulty increases from the second task to the third task having rather extensive vocabulary, complex grammar, the length of the text to be read (260 – 300 words), the change of text source (fiction is difficult to comprehend). Another feature that makes the tasks differ in their level of difficulty is the authenticity, as task 3 is an authentic text which requires high language proficiency. It is also noted that discourse type of the texts is the same in two tasks: task 1 and 3. However, it is recommended to have different discourse types as well as domains in order to motivate the students while doing the test and offer a wide range of communication themes.

Thus, according to the CEFR Global scale, the item comprehensible to a learner/user in the reading test corresponds to the level B2+ which is in between of B2 and C1.

Proficient User	C1	*Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can use language flexibly and effectively*

		for social, academic and professional purposes.
Independent User	B2	*Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.*

Table 19 CEFR Global Scale

Having analyzed the tables described above in respect to the reading test of Year 12 examination, our team came to the following conclusions:

- Level A of Latvian Year 12 exam in the English language can correspond to the CEFR level C1;
- Levels B and C of Latvian Year 12 exam in the English language can correspond to the CEFR level B2+ and B2 respectively.

The process of relating Year 12 examination to the CERF is still in the development, and within this process it appears significant to analyze and relate Latvian Year 12 exam in English to the CERF, finding out how the exam levels correspond to the levels described in the CERF.

Conclusions

To sum up, the process of analyzing Year 12 English language examination Reading test and relating it to the CERF included four stages. These are: familiarization, specification, standardization and empirical validation. Our project group has tried to complete all four stages. We have analyzed the content of Year 12 examination 2010 in the English language, reading comprehension, and its relating to the CERF. Although the members of our group had their own understandings and judgments during the discussion, the procedure allowed us to come to the following conclusion: the Year 12 examination reading tasks test students' general reading ability. The students have to complete three tasks on reading comprehension in 50 minutes. The target performance level can correspond to C1 level according to the CERF.

Relation of Year 12 English Language Examination to CEFR (Alma Bernhards, Ineta Egliena, Santaa Strēle-Ivbule, Olga Smirnova)

In the present research we will depict the process of comparing of the Year 12 Listening Test to the CEFR and the conclusions drawn from the process.

Definition of Listening

Listening is a complex process in which the incoming data, an acoustic signal, must be interpreted by a listener using a wide variety of linguistic knowledge (such as phonology, lexis, syntax, semantics, discourse structure, pragmatics and sociolinguistics) and non-linguistic knowledge (knowledge of the topic, the context and general knowledge about the world and how it works). It is believed that the listening comprehension is "an on-going process of constructing an interpretation of what the text is about and then continually modifying that as new information becomes available" (Buck, 2001:247). Thus meaning is not contained within a text, but it is actively constructed by the listener in an active process of inferencing and hypothesis building. When the task is simple and unambiguous, all competent listeners are likely to come to the same understanding. However, when the comprehension in detail is examined often considerable differences between listener interpretations of many texts can be observed; and more complex and ambiguous the text, the more likely that interpretations will vary (Buck 2001:30).

As a general strategy, Buck (2001:249) advises developing listening tests that concentrate on those characteristics that are unique to listening. The first step in test construction is to define the construct on a conceptual level, and then to operationalise which means selection of texts and development of the test items - part of the test that requires a scorable response from the test-taker (ibid). Buck stresses the idea that operationalising the construct there are two dangers: not covering the whole of the construct – construct-underrepresentation – and including things that do not belong in the construct – construct irrelevant variance (Buck 2001:94). He adds, due to the fact that listening comprehension cannot be examined directly, it is always necessary to give the test-taker some task, and then make inferences about the listener's comprehension based on performance on that task (ibid.). Skills besides listening will always be involved in task performance, and there is always a possibility of construct-irrelevant variance affecting listening test scores.

Since there are no 'hard-and-fast rules' about what an appropriate listening construct is, Buck offers what he calls 'default listening construct' which goes beyond basic linguistic competence, but avoids those aspects of language use that are difficult to assess. It includes: grammatical knowledge, discourse knowledge, and covers almost everything that is unique about listening knowledge:

- the ability to process extended samples of realistic spoken language, automatically and in real time;
- to understand the linguistic information that is unequivocally included in the text;
- to make whatever inferences are unambiguously implicated by the content of the passage (Buck, 2001:114).

It is important to determine which particular competencies are required for each task, in order to ensure that all the required competences are adequately covered.

Bachman and Palmer (2001) suggest six properties as the bases for quality control in test development: *reliability, construct validity, authenticity, interactiveness, impact and practicality*. Reliability is concerned with how accurately the test measures. Construct validity is the extent to which the test measures the right construct. Authenticity is the extent to which the test takers are perceived to share the characteristics of the target-language use tasks. Interactiveness - the extent to which the test tasks engage the same abilities as the target-language us tasks. The scholars agree that tests are powerful things, especially when the stakes are high therefore test-developers need to do what they can to avoid undesirable impact (Bachman and Palmer 2001:17-34).

Materials and Methods

The relation of the Listening test was started by the familiarisation phase, when all the members of the team received a copy of the Year 12 examination in English 2010. We also shared our experiences in using the CEFR.

The next stage was specification of the examination contents (see Figueras et al 2005) in general and its internal validity in particular. The following materials were used at this stage:

a) Latvian Year 12 English language examination paper for the year 2010, Listening Test, as well as the key to the Examination. The whole examination can be found http://visc.gov.lv/eksameni/vispizgl/uzdevumi/2010/vidussk/12kl_angl_val.pdf; the Common European Framework of Reference for Languages and the Manual for Language Testing; detailed statistical information for the Year 2010 examination.

The table below shows the descriptive statistics of the three last years of examination; numbers do not change much, the mean is 47-51%:

	Nr of test-takers	Mean %	St.Dev %
2008	23528	47	18.34
2009	23652	49	20.32
2010	22638	51	19.1

Table 20 Comparative statistics of Year 12 examination. Listening part

We also examined the documents recording test development and marking stages, test specification, and marking instructions. Here we will depict only those for Listening test, as it is the main focus of this research.

The listening part of the examination consisted of three guided listening tasks. The test-takers demonstrated their ability to do the following: in the first task they had to listen to a talk about the history of perfumes. The title of the text was 'Perfumes'. Statements were given below and each of them contained false information. The students had to underline the wrong word or number and write the correct variant.

In the second task, the test-takers listened to a story 'Hotheaded Iceborers' and marked the given statements as true or false.

The third task with the title 'Music and Art' required listening to an interview about the impact of music and arts lessons on children's progress at school. The test-takers had to choose and circle the correct statement from the four variants.

A standardisation meeting was held after the participants had done all the tasks on their own. Afterwards they shared their impressions together. The results were checked and the difficulty level of the tasks was compared. In some cases the difficulty level of the tasks differed according to the participants, thus, the results were written down and then standardized according to the CEFR, looking for evidence in the language level descriptions. The results of the examination tasks were analyzed statistically and compared.

.No of tasks	3
Integration of skills	Listening
Total test time	30 minutes
Target performance level	B1-B2
Channel	Listening
Purpose	General proficiency

Table 21 General information about the Listening Test

As we can see in the table above, the Year 12 examination listening tasks test general proficiency asking the students to complete listening tasks in the 30 minutes provided. The target performance levels could correspond to levels B1 and B2+.

Test	Listening Comprehension in English		
Target levels in the curriculum: B2.1			
	Task 1	Task 2	Task 3
Item types	Editing	True/ false	Multiple choice

Source	Text book	Text book	Interview
Length (total 30 minutes)	2:35	1:20	4
Authenticity	Modified	Modified	Modified
Discourse type	Descriptive	Descriptive	Descriptive
Domain	Public	Public	Educational
Topic	Perfumes	Nature	Music and Art
Curriculum linkage			
Number of speakers	1	1	2
Pronunciation	Standard BrE	Standard BrE	Standard BrE
Content	Concrete	Concrete	Abstract
Grammar	Simple	Somewhat complex	Rather complex
Vocabulary	Only frequent	Mostly frequent	Rather extensive
Nr of listening	2	2	2
Input text comprehensibility at level	B1	B1+	B2+

Table 22 Specifications of the Listening Test

Table 6 shows that the first task corresponds to B1 language level. According to the CEFR, the test taker can 'understand straightforward factual information about common everyday [...] related topics, identifying both general messages and specific details; [...] can understand the main points of clear standard speech on familiar matters encountered in [...] work, school, leisure [...]' (CEFR, 2001: 66). However, the task required listening to and recognizing details. It seemed misleading for the participants of the research who performed the tasks themselves.

The second task corresponds to the language level B1+. According to the CEFR, the description of the level between B1 and B2 would be as follows: the test-taker can 'understand simple technical information; [...] can understand announcements and messages on concrete and abstract topics spoken in standard dialect at normal speech' (CEFR, 2001: 67). In this task the students needed to scan and comprehend the information while listening. Nevertheless, it was two-folded: the students not only had to listen and decide whether it was true or false, but also read the statements, which at times sounded confusing. Thus, the students had to do a lot of guessing. Item 14 mentioned "Arctic" to be marked as true/false, but the recording mentioned "Antarctic" so quite a number of experienced students and teachers were confused.

The difficulty level of the third task increased as the text was quite long. It was a modified interview, thus, at some points it seemed as a natural conversation. It was the most difficult listening task where precise understanding was tested. In addition, it was complicated to concentrate on four multiple choice options to each question at the same time and choose the correct one. Even though, the official examination centre claimed that this task corresponded to the C1 language level, during the research project we came to a conclusion that it corresponded to B2+. According to the CEFR, the test-taker can 'keep up with an animated conversation between native speakers' (CEFR, 2001: 66). Moreover, when comparing the description between the levels B2 and C1 in order to gain B2+, in the CEFR it is stated that the student can 'follow complex interactions between third parties [...], even on abstract, complex, unfamiliar topics' (CEFR, 2001: 66), which seemed to be the case with this listening task about 'Music and Art'.

Conclusions

The listening comprehension is "an on-going process of constructing an interpretation of what the text is about and then continually modifying that as new information becomes available".

Thus meaning is not contained within a text, but it is actively constructed by the listener in an active process of inferencing and hypothesis building. (Buck, 2001)

The Year 12 examination listening tasks tested general proficiency asking the students to complete listening tasks in the 30 minutes provided. The target performance levels can correspond to B1 and B2+.

Finally, the year 12 exam listening part 2010 was successful concerning items' difficulty and ability to discriminate individuals who scored high on the test as a whole and individuals who scored low on the test as a whole. Nevertheless, there is still some room for improvement – the items should test listening for gist or details more than listening for precise information.

Relation of Language Use test in Year 12 Examination to the CEFR (Anna Lavrecka)

This article discusses the place of grammar in Latvian state examination and in the CEFR (Common European Framework of Reference for Languages), mainly referring to the CEFR and the examination paper developed by the specialists of State Centre of Education Content of Latviain 2010. In addition, the notion of "language use" will be explored and the viewpoints of various scholars will be presented. The aim of the article is to come to a conclusion whether language use should be included into the state examination of English in future.

The language use part is to be completed in 40 minutes (to do the tasks and later put the answers on a separate answer sheet). The language use part comes third, students do not have breaks between the parts.

The language use part in 2010 consisted of three tasks:

- Task 1 (18 points) was an authentic text (extract from J.K. Rowling's "Harry Potter and the Philosopher's Stone", fiction) with a fill-in-the-blanks task; the answers in the form of multiple choice were provided below the text; the students were to choose the word out of the suggested four that best suited each space, one example was provided. The items are mainly based on grammar and vocabulary;
- Task 2 (10 points) consisted of 10 lines of a connected text (article about the history of Barbie doll from Wikipedia); some of the lines were correct, but in some of them there was an unnecessary word; the students were to put a tick (the line was correct) or a dash (the line was incorrect) and cross out the extra word if it was in the line, two examples were provided; the task deals mainly with grammar aspects;
- Task 3 (17 points) is an open cloze based on a fictional text (Ch. Darwin "The Voyage of the Beagle"); one example was provided; the task deals with the language at discourse level, correcting minor faults in the text that do not influence its general understanding.

Having explored the tasks, the author came to the conclusion that language use part of the Latvian state examination of the English language generally deals with two language aspects: grammar and vocabulary.

After Latvian state examination in English has been described, it is necessary to discuss the term "language use". Needless to say, the opinions on this issue vary noticeably. Harmer (2006) views

language in use as an issue completely separate from the language skills and language aspect. He refers to four characteristics of this notion:

- purpose – the aim the speakers strives to achieve or, in other words, language functions realised through the language (e.g., invite, apologise, etc.);
- appropriacy – making use of the factors influencing the choice of language (setting, participants, gender, channel, topic) and selecting the language accordingly;
- language as discourse – to use language elements in discourse ("language used in context over an extended period" – Harmer, 2006: 25), namely situation, problem, response and evaluation (Aston, 1997 in Harmer, 2006); e.g., applying the patterns a typical conversation in a certain situation may follow;
- genre – making use of various types of discourse; e.g., announcement, advertisement, and alike.

According to Harmer (2006), language use means rather the practical application of the target language at discourse level, taking into account all the factors that may influence it than, for example, knowledge of its grammatical rules and regularities.

Batstone (1994) also points out that in real-life language situations there is no actual time for conscious application of grammar rules, which leads him to the conclusion about the interdependence of language use and procedural knowledge of the language(any language pattern or routine used repeatedly). This means that language use deals generally with the production of the appropriate and relevant language discourse rather than the attention is paid to the separate language items and their combination patterns.

Similar point of view is expressed by Brumfit (1984), who sees language use as "a process of approximating the public avowals we make of our perceptions to other people's public avowals, to the extent necessary for us to perform effectively whatever it is we want to do with other people, or to obtain whatever it is we want to obtain from other people" (Brumfit, 1994: 29). Brumfit's understanding of language use presumes that language use requires not only the knowledge of grammar and vocabulary of the target language, but also many more issues of language and communication, such as register, development of the interpersonal intelligence, pragmatic strategies, and alike.

Moreover, Widdowson (1978) distinguishes between language usage (language rules) and language use (the ability to apply the knowledge of these rules in order to reach the effectiveness of the communication).

In sum, the term "language use" refers to the language produced in order to reach the communicative aim in the most efficient way possible; in other words, to get the message across effectively. It does not mean that syntactical or lexical aspects are sacrificed; rather, it implies the interaction of various language and communicative factors.

The ideas presented above weakly correspond to the tasks of Latvian state examination. The problem lies not in the length of discourse or the amount of student input, but in the purpose of the tasks which requires the students to reveal the quality of their performance in grammar and vocabulary of the target language. Since in most cases this part of the examination deals with grammar and vocabulary issues, I suggest changing the term "language use" to the term "language accuracy" which, according to Brumfit (1984), means producing examples of language according to certain requirements (phonological, syntactical, lexical, functional or stylistic) and implies the focus on the target language rather than on the message conveyed by the language user.

After the terminology has been discussed, there are other confusing issues regarding the relation of Latvian state examination of the English language to the CEFR standards. Firstly, the CEFR standards deal with four language skills but do not mention language use or language accuracy as a separate competence to be assessed in the examinations of foreign languages. Secondly, if Latvia decides to keep "language use" as a part of state examination in English, the strategy for relating the result to the CEFR language proficiency levels is necessary to be developed.

It is notable that the CEFR does not define the term "language use", but uses categories for its description. These categories are as follows:

- the context of language use (domains, situations, conditions or constraints, user's mental context, interlocutor's mental context);
- communication themes (e.g., travelling, work);
- communicative tasks and purposes (e.g., communication at a workplace);
- communicative language processes (e.g., planning, formulating);
- texts (e.g., texts, genres) (CEFR, pp. 43-100).

The categories of description presented above are in correspondence with the scholars' viewpoints (Batstone, Brumfit, Harmer, Widdowson) on language use and, again, in contrast with the tasks in language use section in Latvian state examination of the English language.

Furthermore, language use is not mentioned in the description of general CEFR language proficiency levels; these levels are based on the development of communicative competence, as well as on the languages users' abilities rather than on their faults. There are accuracy issues mentioned in language skills assessment grids as criterion; for example, *Table C4: Written Assessment Criteria Grid* where grammatical structures are being discussed according to language proficiency levels (CEFR, p. 199). Also, the notions of *grammatical accuracy* and *vocabulary control* are included in the aspects of communicative language competence in CEFR examination relating manual. However, in reference to current tasks offered by the language use section in Latvian state examination, there are two descriptions of language proficiency levels considering the lexical competence ("knowledge of, and ability to use, the vocabulary of a language, consists of lexical elements and grammatical elements" –CEFR, p. 110) and grammatical competence ("knowledge of, and ability to use, the grammatical resources of a language; grammar is seen as the set of principles governing the assembly of elements into meaningful labelled and bracketed strings (sentences). Grammatical competence is the ability to understand and express meaning by producing and recognising well-formed phrases and sentences in accordance with these principles (as opposed to memorising and reproducing them as fixed formulae)." – CEFR, p. 112) of the language users.

GENERAL LINGUISTIC RANGE

C1

Can select an appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say. Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.

B2

Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so. Has a sufficient range of language to describe unpredictable situations,

explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.

B1

Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times. Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words.

GRAMMATICAL ACCURACY

C1

Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot. Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.

B2

Shows a relatively high degree of grammatical control. Does not make mistakes which lead to

misunderstanding. Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express.

B1

Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.

Having studied the CEFR standards and compared the language use section of Latvian state examination of the English language in the 12th Grade, I have come to the following conclusions. Firstly, the notions of language use in Latvian state examination and in CEFR standards differ dramatically. While Latvian examination paper requires the performance in grammar and lexis, CEFR standards view language use as the complex of various skills and competencies necessary for the effective communication process. Therefore, the term used by Latvian specialists should be changed to a more appropriate one.

	Constructs	CEFR	Year 12 exam
Accuracy	x	x	x
Fluency	x	x	
Pragmatic strategies	x		
Appropriacy	x		x

Function	x	x
Genre	x	x
Context	x	x
Language processes	x	x
Interperson communication	x	

Table 23 Relation of the language Use paper to CEFR and State curriculum

Secondly, CEFR standards do not require assessing language users' language use (or accuracy) separately. It is involved in the assessment of language skills' development as a criterion for the assessment. Moreover, it is, to my mind, unreasonable to assess language use as such outside the context; it should be assessed within the relevant discourse as a tool for the fulfilment of the communicative purpose. Consequently, if Latvia is interested in relating state examination to CEFR standards, it is advisable for the language use section to be removed from Latvian state examination of the English language.

Relation of the Latvian Year 12 English Writing Test 2010 (Zilgme Eglīte, Aira Misa)

The authors have made a survey of the theories of writing assessment and evaluated Latvian Year 12 Writing Test in English. The ways in which the Writing Test is linked to the Common European Framework of Reference for Languages (CEFR) is explored by describing the level B2, being achieved by half of the students. The aim was to see the correspondence of demands as set by the Education Content and Examination Centre (VISC) of Latvia, as well as to evaluate the validity and reliability of the test.

The authors were standardised according to the procedures of standardisation for examination markers, and then they marked 15 tasks from the examination and compared their markings to the official results. In addition, the statistical data on the exam results from the VISC website were evaluated.

Theoretical Background of the Writing test

"Writing is a complex activity involving thinking, planning and organizing; in addition, it requires from the writer the knowledge of spelling and punctuation, the so called orthography" (Hamp-Lyons, 2003: 165). This is when one writes in his/ her native language. Writing in a foreign language is much more demanding, as a writer needs not only the mentioned skills, but also linguistic knowledge, discourse knowledge and sociolinguistic knowledge of the target language (Weigle, 2002: 29). To give examples of a few of these, the linguistic knowledge involves sufficient vocabulary and mastery of grammar, the discourse knowledge involves knowledge of cohesion devices and structures of texts of different genres, and the sociolinguistic knowledge includes considering your audience and choosing to write in an appropriate degree of formality (ibid., 30).

Theory on assessing writing

The writing assessment can take two forms: direct and indirect. The indirect writing assessment is by multiple-choice, grammar completion etc. and is easy, fast, and reliable to evaluate (Hamp-Lyons, 162). The direct writing assessment is an assessment of a concrete performance of

writing, of a concrete writing task. Because of the complexity of writing, a person might write differently on different occasions and on different topics It requires more skill and time to evaluate them, and the score given is less reliable. (Hamp-Lyons, 2003: 165). The writing part of the Year 12 Exam in English implies the direct writing assessment.

Reliability

According to the language assessment experts, "Reliability is an essential test quality that can be thought of as the degree to which test scores are free from measurement error. In a language test, any factor other than the ability being measured that affects the test score is a potential source of measurement error" (Bachman, Davidson, Ryan, Chol,1995: 52). Reliability is also a measurement feature: the ability to repeatedly deliver the same results (Rasinger, 2010: 55). That is, when asked to write an essay, a student would get the same score.

The writing scores are reliable only for 80 % or less. A good score reliability (of 75% and more) can be achieved by having two (or more) raters agree on a score, plus validity. Still, as a person can write the pieces of different quality, only a concrete performance can be rated.

There are two ways two improve reliability – the "Devon method", where writing is marked be several examiners, and "moderation" system, where accuracy of raters is tested by sampling their score. The first, however, is more reliable (Hamp-Lyons, 2003: 163-164). The Latvian Year 12 English exam is marked by two markers, and in case of the discrepancy, is passed over to a marking commission.

According to American Psychological Association, "Validity is the most important quality of test interpretation or use, it is the extent to which the inferences or decisions we make on the basis of test scores are meaningful, appropriate, and useful (1985, cited in Bachman, 2001: 25).

Validity is also a measurement feature but concerned with whether we are measuring what we are supposed to measure (Rasinger, 2010: 26). In writing assessment, traditionally four types of validity have been considered, according to Hamp-Lyons:

- face validity (whether the test appears valid to an outsider);
- content validity (grounded in some evidence, as e.g. the relevance of the content to the test taker);
- criterion validity (based on correlation between the test and other measures);
- and construct validity.

The construct validity is the most topical of them and consists of the following facets that all have to be 'valid': the task, the writer, the scoring procedure, the reader and the text itself.
Concerning the validity of the task, it is important to consider, whether the prompt (input text) gives sufficient information to which to respond. Some authors argue, that there should be a choice of prompts, not only to help the writer who might have nothing to say on a particular topic, but also, because it is believed that 'a choice of prompts is likely to help students'; having a choice, relieves anxiety and the test takers feel more confident (Hamp-Lyons, 2003: 172). The writer validity consists of assigning a topic that is relevant for the group of students to be tested concerning their age, experience and interests. Also, the time allocated for the completion of the task should be reasonable.

There are three categories of the scoring procedure:

1) holistic scoring, which means giving a score for the general impression of the quality of writing, compared to others tested at the same time. This procedure is the most unreliable one (Hamp-Lyons, 2003: 175);
2) primary trait scoring means that one aspect of writing is selected as primary for judging its quality;
3) Multiple trait scoring assesses multiple aspects of writing, taking into account the complexity of the skill; one trait can have a higher score than the other. This category of scoring procedure is used in Latvian Year 12 English exam.

Finally, for the reader (also called a rater, a marker, an assessor, an examiner) to be '*valid*', he/she has to be trained ('*standardised*') to minimise the subjectivity and to increase the reliability and validity. Newcomb in 1977 established that race, sex, geographic origin can affect the raters' responses to essays, and so can, presumably, age, ethnic origin, cultural context and raters' own experience of learning, test taking, as well as teaching. The same has been found out to be true about the expert and novice raters: experienced raters pay more attention to 'higher-order aspects of writing', while novice raters – to a 'lower-order aspects'. It is claimed that despite training and carefully defined criteria, raters 'act as individuals, using their own values' (Hump-Lyon, 2003: 178-179).

After Year 12 Writing exam study in 2007 it was concluded that the target curriculum level should be C1 (Kalnbērziņa, 2007: 8). According to the guidelines issued by the Education Content and Examination Centre (ISEC), the Standard of secondary education encourages students to reach the level B2 to C1 (ISEC, 2010).

The exam level is in the range from B1 to C1. As, according to the research in 2007, the Latvian Year 12 levels B and C correspond to CEFR level B2, and in 2010 49 % of the test takers reached this level (see Figure 1), we are going to describe this level in detail.

B2 is also called *Vantage* level – a level at which the user of a foreign language is able to provide 'adequate response to situations normally encountered' (Trim, cited in CEFR: 23). It corresponds to the level tested by Cambridge ESOL FCE, or IELTS band of 6 (Kalnbērziņa, 2007: 8).

According to CEFR, a B2 user can write in all kinds of situations and contexts. CEFR provides the following scale descriptions for the level. In overall written production, a B2 means, that a person can 'write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources' (CEFR: 61). B2 is not as demanding level as C1, as a C1 user can write on complex subjects, and support arguments with a more complex structure. But it asks more then B1 where one is required to join shorter elements into 'a linear sequence' (ibid.).

A B2 level person can write 'a review of a film, book or play', even descriptions of 'imaginary experiences'. In the field of creative writing, they can follow the genre conventions. What concerns reports and essays, then B2 users can give arguments for and against, as well as synthesise information (CEFR: 62). A B2 person can take part in written interactions, take notes, fill out forms, and write messages and correspondence. For example, such a person can 'write letters conveying degrees of emotion and highlighting the personal significance of events' (CEFR: 83).A B2 person possesses an adequate orthographic competence and 'follows standard layout and paragraphing conventions', and is reasonably accurate in spelling and punctuation (CEFR: 118).

The Description of Year 12 (2010) Writing Part

The Writing Test consisted of **three tasks** in total, with the maximum of 60 points to be received. The timeline for all three tasks was 80 minutes.

Task 1 required writing a paragraph with arguments for and against computer games; task 2 was a description of a sightseeing trip for a brochure with three specifications to be covered; and, finally, task 3 was a composition on the health with a description and explanation of its three main factors, along with the student's own suggestions.

Task specifics	**Task 1**	**Task 2**	**Task 3**
Language	English	English	English
Language level	B1	B2	B2
Time for task	Not specified	Not specified	Not specified
Length of task	80 words	110 words	250 words
Content	Specified	Specified	Specified
Discourse type	Argumentative	Descriptive	Argumentative
Type of input	Textual	Textual	Textual
Topic	Computer games	Sightseeing	Health

Table 24 Analysis of each task according to the Common European Framework (2010)

The three tasks were evaluated in accordance with the CEFR for Languages assessment as to the contents, the organization, the grammar, the vocabulary, and the orthography. The scale ranged from 0 (not enough to evaluate) up to 5: "Can produce clear, well-structured, detailed text on complex subjects" (CEFR level: C1). The highest possible CEFR Level attained for the task 1 was level B2: "Can produce clear, detailed text on a wide range of subjects", for the task 2 that was the CEFR level C1: "Can produce clear, well-structured, detailed text on complex subjects", and for the task 3 the highest level attained was the CEFR level C1 again.

The test included two types of **text discourse**: the argumentative in task 1 and in the task 3, and the descriptive in task 2. The **domains of texts** were quite varied, ranging from public (e.g., computer games in task 1; sightseeing, and tourism in task 2) to educational (e.g., the health in task 3).

The Common European Framework's demands do not fully correspond to the demands of the Latvian test developers. For example, the CEFR does not require that the criterion for the number of words is observed, in contrary to the Latvian Year 12 Writing test in English. However, the VISC of Latvia suggests reaching the CEFR Level of B2 in the Writing test: "Can produce clear, detailed text on a wide range of subjects".

The test assessment was made, firstly, for the coherence in writing, i.e., the overall organization of the text.

Response	**Task 1**	**Task 2**	**Task 3**
No of words expected	60 words	110 words	250 words
Register	Neutral	Neutral	Neutral

Domain	Public	Public	Educational
Grammar	Simple structures	Frequent errors	Rather high grammatical control
Vocabulary	Frequent, basic	Extended	Extended
Cognitive processing	Reproduction	Knowledge transformation	Knowledge transformation
Content knowledge	Responds the task	Satisfactory	Responds the task

Table 25 The Response Analysis in the Year 12 Writing Test (2010)

Evaluation of Year 12 Exam (2010) Writing Part

We consider Latvian Year 12 Writing part of the English exam reliable and valid (face and content validity), as well as thoroughly related to the CEFR (criterion validity). The tasks were well organised, and the prompts were appropriate. Task 1 and Task 3 were somewhat similar in their discourse type – being argumentative, and it would probably have been better to expect a different discourse type in each of the three tasks. The discourse types can be a description, a narration, a commentary, an exposition and others, listed under the 'Functional competence' in CEFR (CEFR: 126). This problem seems to be avoided in the forthcoming exam in 2011, as only two tasks are planned, with clearly defined functional differences: a piece of written interaction (a letter, a recommendation), and an argumentative essay (VISC, 2010). We hope that the tasks will conform to what Hamp-Lyons calls the 'writer validity', and will be on a topic familiar to the school leavers, as, for example, a letter of recommendation might not be suitable for the purpose.

We would also like to add our thoughts on the validity of the scoring procedure. Our class of MA students were introduced to the exam marking standards, and we all marked 15 writing tasks from the Latvian Year 12 Exam (2010): 5 writings of task 1, 5 of task 2, and 5 of task 3. An example of our markings for task 3 for five different test taker papers is illustrated in the Table 3 below.

Task 3	Student Marker 1	Student Marker 2	Student Marker 3	Student Marker 4	Student Marker 4	Student Marker 5	Student Marker 6	Student Marker 7	Student Marker 8	Student Marker Average	Official Maring
Paper 1	11	8	13	7	11	8	8	8	8	**9,1**	**5**
Paper 2	22	21	23	22	22	22	21	22	22	**22**	**19**
Paper 3	12	9	14	11	16	12	17	13	13	**13**	**12**
Paper 4	7	6	7	7	12	9	11	8	8	**8,5**	**0**

Paper 5	7	7	7	7	8	14	6	8	6	**7,8**	**8**

Table 26 The Standardized Evaluation of the Year 12 Writing Test (2010)

In general, our markings were slightly more lenient then the official ones, but that might be due to our lack of experience. However, we were concerned about the discrepancy between our marking of paper 4 (highlighted in the table), where our average was 8,5, but the official marking was 0. It turned out, that that the markers are instructed to give 0 points if the word count in the paper is considerably below the required, and 0 is given to all criteria (contents, organisation, grammar, vocabulary and orthography. The CEFR is not explicit on the length of the writing task as the indicator of its level, and we think that if the task is too short, a mark can still be given for some of the criteria, even if one of the criteria receives a '0'. The test specification requires giving a '0' for a criterion if there is not enough written to evaluate, and also marking instructions of 2009, published on the website of VISC, give examples of such a marking method applied (VISC, 2009: 12).

Our concern is confirmed by the statistical diagram on the assessment of the writing skill (see figure 2). In general, the histogram shows well spread examination results, with the mean (central tendency) of 47,06 %, the mode around 60 % and standard deviation of 23 %. But the first two bars of the histogram show a tendency that might indicate a problem with the validity of scoring procedure. In an ideal test histogram, the first bars are the lowest, close to 0 – there should be very few who fail the test if they have been preparing for it for years. This histogram shows that about 1200 test takers failed the writing part, and about 2000 test takers, or 10 % of the total number received 10% or less of the score. We do not have the evidence for it, but this might have been due to the markers automatically assigning '0' to every paper with the word count below (or above) the optimal.

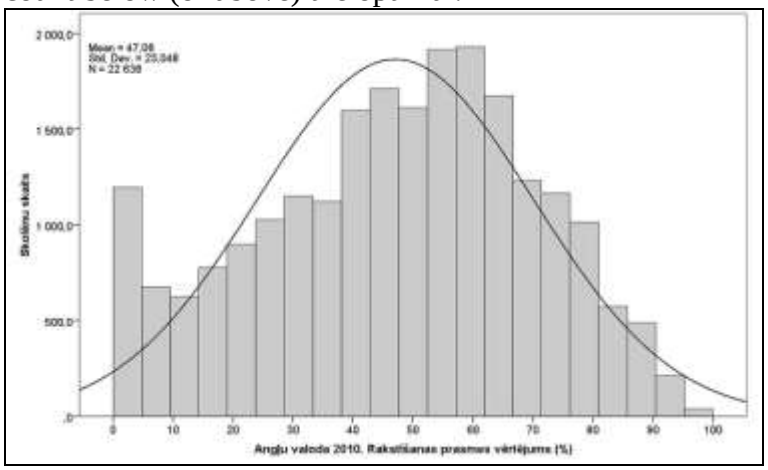


Figure 11 English 2010 Assessment of the Writing Skill (available at www.visc.gov.lv**)**

Apart from the study of statistical histogram and our replication of the marking procedure, another way to assess the reliability of scores is by looking at correlations. Markings by the first and the second rater can be correlated, but also different skill results can be correlated against each other and against the overall test result (total performance). A visual comparison of histograms for all skills and the total performance show that in no other skill have there been so many (over 1000) zero marks. For example, in reading, another 'textual' skill, less then 60 students received a '0'. Speaking, which is another 'productive' skill, also had a high number of

‘0’ results – about 500, but it is still two times less than the number of those who failed in the writing test.

Whatever the reason for the high number of failing scores in this skill, it deserves attention of the exam and marking system designers, as ‘the test results affect students’ lives, sometimes critically, opening or closing the opportunities on which their entire futures depend’ (Hamp-Lyons, 2003: 182).

Conclusions

Our research shows that the Writing part of the Year 12 English Exam is well developed, its marking scale is carefully related to the level scales of the Common European Framework of Reference for Languages, and the overall test is reliable and valid. An issue still deserving attention is finding the reason for a high number of test takers achieving only zero result in the writing part, which might be due to prioritising the word count over other criteria that can be evaluated.

Relation of the Year 12 English Language Examination Speaking Test to CEFR (Tatjana Savenkova, Margarita Šendo, Žanna Moskovkina)

Since the advent of Communicative Language Teaching Approach (Richards and Rodgers, 2001: 151) the development of communicative competence became central to foreign language teaching. Moreover, according to the Standard of Secondary Education of the Subject of Foreign Language of Latvia communicative and language competence is included in the obligatory content of the subject. Thus, speaking skills are central to the curriculum in language teaching. Fulcher (2003: 23) defines speaking as ‘the verbal use of language to communicate with others’. Lado (1965: 239) once remarked that ‘the ability to speak a foreign language is without doubt the most highly prized language skill and rightly so’. Therefore, measuring speaking skills has become central to foreign language testing. However, Fulcher (2003:1) states that the theory and the practice of testing second language speaking is the youngest field of language testing. Luoma (2004: 1) claims that assessing speaking is challenging as many factors influence one’s impression of how well someone can speak a language. Furthermore, testing these skills is a difficult task due to the complex nature of speaking (O’Sullivan, 2008: Online 2).

It is obvious that speaking is the most difficult skill to test due to its complexity. Luoma (2004: 8-28) points out the following features of the spoken language:

- *The sound of speech is meaningful* and therefore, is a thorny issue for language assessment’. It helps people judge the speaker, in other words, “people use their speech to create an image of themselves to others’.
- *Composed of idea units* (short phrases and clauses). Moreover, ‘the grammar of these strings of idea units is simpler than that of the written language’.
- *May be planned or unplanned.* Ochs (1979, cited in Luoma: 2004:12) states that speeches, lectures, conference presentations, and expert discussions involve planned speech ‘where the speakers have prepared and possibly rehearsed their presentations in advance’. Unplanned speech, in contrast, ‘is spoken on the spur of the moment, often in reaction to other speakers’. Thus, planned speech tends to be more formal, whereas, unplanned speech can range from formal to informal. Therefore, unplanned speech has simpler grammar and consists of short phrases and short turns between speakers.
- *Employs fixed phrases, fillers and hesitation markers,* since they are typical for spoken-language.

- *The internal structure of idea units.* Topicalisation, which 'breaks the standard word order' and tails, which 'are noun phrases that come at the end of a clause' are typical for spoken language when the speaker wants to emphasize the topic.
- *Slips and errors* are typical for normal speech (e.g. mispronounced words mixed sounds and wrong words due to inattention). Despite the fact that native speakers also tend to have slips and errors, 'in the speech of second or foreign language learners these mysteriously acquire special significance'. Furthermore, some errors are typical both for native and non-native speakers, whereas, there are some errors, which are typical only for language learners. For this reason, 'assessment designers may have to provide special training to raters to help them outgrow a possible tendency to count each 'error' that they hear'.
- *Involved reciprocity.* By reciprocity Bygate (1987, cited in Luoma: 2004: 20) 'means that speakers react to each other and take turns to produce the text of their speech together'. This is how speakers process demands of speech, but it also has a social dimension 'in that their phrases and turn-taking patterns create and reflect the social relationship between them'.
- *Shows variation.* Firstly, people talk to each other for different purposes. Brown *et al.* (1984, cited in Luoma, 2004: 22) characterises to extremes: 'chatting or listener related talk, and information-related talk. … Moreover, both types of talk can occur in one and the same speech event; in fact, this is what normally happens'. Secondly, the social and situaional context in which the talk happens also influences what gets said. Hymes (1972, 247:248) suggests the acronym SPEAKING to present the following components of speech situation: *Settings* - setting, scene; *Participants* - speaker/hearer, *Ends* - functions (transactional or interactional) and outcomes (effects); *Act sequences* - message form and content; *Key* – tone, mood or manner; *Instrumentalities* – channel (e. g. verbal, non-verbal, face-to-face, written, etc.) and code (language variety); *Norms* - norms of interaction and interpretation; *Genres* – genre (e.g., lecture, seminar, story). Thus, assessment developers may find this framework useful, since 'it will help them describe the test construct in some detail' (Luoma, 2004: 25). Moreover, the framework can serve as a good guide for 'the comparison of individual test administrations against each other, which is important for fairness' (ibid.). Thirdly, speaker's choice of words determines speaker roles and role relationships. Thus, the speaker roles and role relationships depend not only on social and situational context, but also on 'the way that politeness appears in the talk'. Thus, due to politeness people do not communicate 'maximally efficiently'. However, communication would be efficient if they followed Grice's (1975, 45-46) four conversational maxims:
- 'Quantity: make your contribution as informative as is required (for the current purposes of the exchange); do not make your contribution more informative than is required.
- Quality: try to make your contribution one that is true; do not say what you believe to be false; do not say that for which you lack adequate evidence.
- Relation: be relevant.
- Manner: be perspicuous; avoid obscurity of expression; avoid ambiguity; be brief; be orderly'.

Thus, politeness is a very important feature that has a relative power since it influences speaker role and role relationships. Moreover, politeness is a difficult concept for assessment since 'it is guided by principles rather than roles….[Moreover] it is interpersonal and social, and the social relationship between test participants is artificial'. Luoma (2004:27) suggests assessing politeness 'in gross terms only, for instance on the three levels of appropriate, somewhat appropriate and questionable or worse'.

To sum up, speaking is meaningful interaction between people. Moreover, its nature is complex and therefore, it is difficult to assess. The difficulty arises from the fact that grammar

and vocabulary of spoken language differ from those of written language. Moreover, the choice of words depends on the purpose of the talk, social and situational context. Therefore, it is necessary to analyse the kind of speaking that needs to be assessed 'in a particular context in terms of social and situational need' (Louma, 2004: 27-28). Furthermore, it is necessary to remember that 'speaking is interactive' when designing 'rating criteria and procedures, and reward examinees when they repeat or mirror the other speaker's phrases and structures or develop topics by referring to earlier turns and building on them, because this shows that they know how to work interactively with other speakers' (ibid.). Considering mentioned-above, the developers of speaking assessments must be clear about what is speaking and then Lauma (2004:28) proposes the following:

- 'define the kind of speaking they want to test in a particular context;
- develop tasks and rating criteria that tests this;
- inform the examinees about what they test;
- and make sure that the testing and rating processes actually follow the stated plans'.

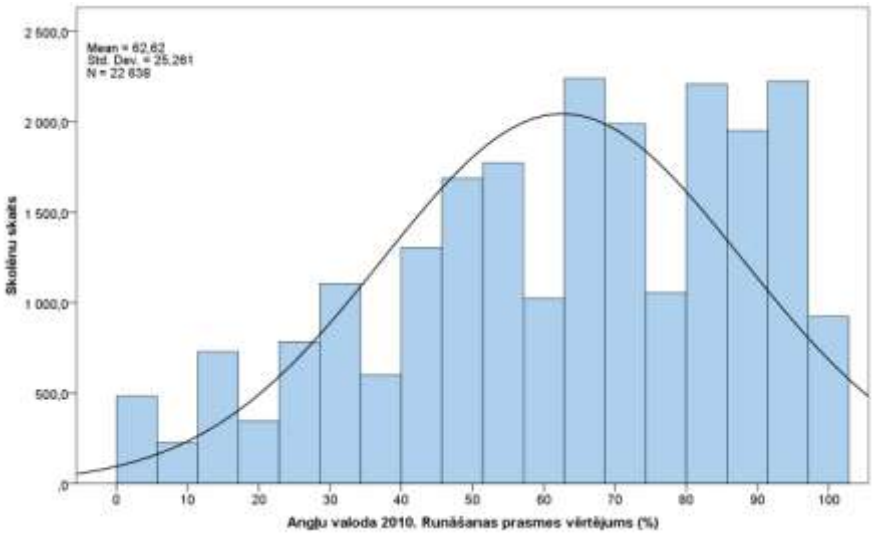
Since the Common European Framework of Reference offers a framework, which has a growing role for language testers; it is necessary to set the relationship between Latvian Year 12 exam and speaking test in particular and the CEFR. In order to do relation the linking process was based on sets of procedures has to be applied (Figueras et al, 2009) The linking process is proposed by the Council of Europe's Manual Relating Language Examination to the CEFR (ibid.)

Table 2 provides the descriptive statistics of the speaking part of Latvian Year 12 examination.

	Nr. of test-takers	Mean %	St. Dev %
2008	23526	59.4467	24.2031
2009	23652	61.2289612	25.000466085
2010	22638	62.62	25.261

Table 27 Comparative statistics of the speaking part of Latvian Year 12 examination

As we can see from Table 27, the descriptive statistics of the examination does not change much from one year to another; the mean is 47-51 % in spite of ever-changing numbers of test takers. Standard deviation shows a good spread of results.



It is obvious from Figure 2 that the number of results above the mean is higher and, therefore, tasks did not cause great difficulty for the examinees.

Figure 12 Speaking skill assessment. Year 12 Exam in English 2010. (www.visc.gov.lv)

Latvian Year 12 speaking test specification:

The speaking test consists of three guided speaking tasks. All three tasks are not connected among themselves. In Task one the examinee is required to answer one or two questions. Task one is a warm up, therefore, the examinee's answers are not evaluated. In task three examinees are required to give their opinion on the given topic and on a number of questions. Examinee's papers do not include the questions; the questions are included in the examiner's paper. They have one minute to prepare their speech. Task three is a role-play. Examinee's paper includes key points upon which they have to build a dialogue with the examiner. They have one minute to prepare their speech. At the beginning of each task, the examiner reads out instruction for the examinee. The exam is computer - recorded and assessed by the assessor. The rating method of all three tasks is a descriptive scale which includes the following criteria: communication strategies and interaction, task achievement, accuracy, fluency, pronunciation. Each criterion is evaluated from one to six points.

According to State Education Content Centre of Latvia considerable changes have been made to 2011 Year 12 speaking test. The specification of Latvian Year 12 speaking test 2011 is adapted from State Education Content Centre of Latvia. The number of tasks has not been changed and The exam includes three communicative tasks: interview, dialogue, and monologue.

Task 1	Interview *(questions)*	3–5 min
Task 2	Dialogue *(role-play)*	3–5 min
Task 3	Monologue *(summary, opinion about 60 to 80 word long text)*	5 min

Table 28 Year 12 Exam in English. Speaking 2011

According to Table 13, 2011 Year 12 speaking test includes 3 tasks which are not interconnected. Task 1 is an interview. The examinee chooses a paper and the examiner states the topic of the chosen paper. The examinee is required to answer five questions on the given topic. These questions are not included in the examinee's paper. The examinee has 3 -5 minutes to answer five questions. These tasks aim at assessing the examinee's ability to answer the questions spontaneously, express and support his/her ideas. The rating method of all three tasks is a descriptive scale which includes the following criteria: vocabulary, grammar, fluency, pronunciation and intonation.

The questions concern the topics provided in Table 14.

Number	**Topic**
1.	Personal information
2.	Home and place of residence.
3.	Daily routine
4.	Leisure and entertainment
5.	Travelling
6.	Relationship
7.	Health and body
8.	Education

9.	Shopping
10	Food and drinks
11.	Services
12.	Places
13.	Language
14.	Weather

Table 29 Communicative topics

Task two is a role-play. This task is not changed and has got the same structure and features as role-plays in the previous year. The examinee's paper includes key points upon which they have to build a dialogue with the examiner. One minute preparation time is given. The role-play is 3-5 minutes long and aims at assessing the examinee's ability to communicate in a given situation.

Task three is a summary and opinion on a 60 - 80 words long text. The examinee has two minutes to read the text and prepare the answer. No note-taking is allowed. The present task aims at accessing the examinee's ability to produce a coherent and cohesive monologue in which they have to summarise the text; express and support their opinion. During the monologue the examiner is not allowed to interrupt the examinee.

Table five is the next stage of specification of Year 12 examination contents in terms of the CEFR Grid for speaking, developed by ALTE members (Online 6).

General information about the speaking test		
1.	Name of test	Latvian Year 12 examination Speaking test
	Component	Speaking component
2.	Target language	English
3.	No of tasks in the speaking component	3
4.	Integration of skills	Speaking, listening
5.	Total duration of speaking component (including preparation time)	Approximately 15 min (of which 3 minutes - preparation time)
6.	Target performance level CEFR	B1-C1
7.	Channel	Face-to-face (computer-audio)
8.	Test purpose	General proficiency

Table 30 Test specifications

Thus, Table 15 reflects the comparison of Year 12 examination speaking test with the CEFR tests demands. The speaking test assesses general proficiency and asks students to produce face-face tasks during the 15 minutes provided.

Next, it is necessary to compare each task of 2011 Year 12 speaking test to the CEFR Grid. For this reason the CEFR Grid, developed by ALTE members was applied. Table 31 gives the opportunity to analyse each task separately.

	Task 1	Task 2	Task 3
Language of instructions/rubric	English	English	English
Instructions spoken/written (channel)	spoken, written	written	spoken, written
Level of language of instructions/rubrics	Easier that level of test	Same as level of test	Easier that level of test
Task duration	3-5 min	5 min	3-5 min
Number of assessors present	1	1	1
Recorded?	Yes-audio	Yes-audio	Yes-audio
Control/guidance by the task (flexibility of task frame)	partially controlled	partially controlled	partially controlled
Control/guidance by interlocutor (flexibility of interlocutor frame)	partially controlled format	partially controlled format	partially controlled format
Interaction type	Dialogue: candidate/ examiner	Role-play	Monologue
Discourse mode (genre)	Interview	Conversation	Speech, presentation.
Audience (real)	assessor	assessor	assessor
Type of prompt	Textual (written sentence, instructions)	Textual (written sentence, instructions)	Textual (written sentence, instructions)
Setting (imagined)	Social	Social	Educational

Table 31 Analyses of each task in terms of the CEFR

Response	Task 1	Task 2	Task 3
Length of response expected	3-5 min	5 min	3-5 min
Text type	Discourse level	Discourse level	Discourse level
Rhetorical functions	Expressing, giving opinions	Asking for information	Summarising, giving opinion
Register	neutral	neutral	formal
Domain	personal, public	public	educational
Grammatical level	Mainly simple	Limited range	Wide range
Lexical level	Mainly frequent	Extended vocabulary	Wide
Discourse features	limited	competent	competent
Situational authenticity	low	high	low
Interactional authenticity	high	high	high
Cognitive processing	reproduction	knowledge transformation	knowledge transformation
Content knowledge	common	common	wide range
Task purpose	conative	phatic	conative

Table 32 Results of the benchmarking

It is obvious from Table 31 that examinees are partially controlled by the task and by the interlocutor. Examinees are engaged in different types of interaction such as dialogue, role-play, and monologue. All tasks are accompanied by textual prompts.

Table 32 describes the examinees' response in three tasks in terms of the CEFR. The analysis of the response in the speaking test shows that the level of difficulty increases throughout the test of speaking using different variables more typical samples to illustrate performance at a given level both for standardisation training and to serve as a point of reference in making future decisions about performances of candidates.

The next stage is standardisation, which includes benchmarking and standard setting. According to Figueras (et al 2009) 'benchmarking is a tool of standard setting used for assisting markers in giving valid judgements in holistically rated tests'. Benchmarking meeting was held in order to benchmark a sample-speaking test to the levels in the CEFR and to establish agreement and understanding of the newly designed marking scale. A group of benchmarkers was provided with photocopies of speaking skill scale for this particular exam and a scale with levels and their description taken from the CEFR, sample exam papers and sample exam recordings. Sample exam recordings were listened and assessed with the help of the assessment scale designed for this particular test. The benchmarkers were asked to compare the results after having listened to each recording delivering the result they thought to be appropriate for that particular answer. After all the examinees were given points the recordings were listened repeatedly which allowed setting the students performance on the CEFR level.

Paper	**CEFR**	**Rationale**	**Speaking test**
1.	C1	uses languages flexibly and effectively, developing particular points, providing clear descriptions, opinions and arguments; has a good command of vocabulary, consistent grammatical accuracy, a clear, natural-like pronunciation and intonation	29
2.	B2	expresses himself with ease, fluently participating in the given situation, using various language structures and having a sufficient vocabulary to express himself; however, there are some slips in sentence structure.	23
3.	B1	develops point of view having limited precision, which causes difficulty to follow the point being delivered by the speaker; nevertheless grammatical errors do not cause problems in communication	19
4.	A2	having sufficient vocabulary on everyday repertoire, provides concrete information in short contributions, even though pauses and reformulation are evident	15

Table 34 Analyses of the response in the speaking test

As we can see from the table below the student performance in Year 12 speaking test ranges from C1 to A2. According to Kalnberzina (2006:6) Latvian Year 12 examination levels correspond to the CEFR levels in the following way (Table 19).

Latvian Year 12 Level	**The CEFR level**
A	C1
B,C	B2

D	B1
E	A2

Table 34 Latvian Year 12 examination levels correspondence to the CEFR

In order to pilot Latvian Year 12 speaking test rating scale, the team has approbated it by evaluating students according to it. In the result of benchmarking it turned out that the marking scale contained ambiguous phrases that resulted in evidential misunderstanding in marking. Thus, when marking the first recording, the role-play, six benchmarkers agreed on four points, four markers offered three points and two markers gave two points. Similar situation was with the third task, which is a monologue, seven people agreed on three points, three markers agreed on two points and two people gave one point to the speaker. Out of this situation, the following conclusion is to be made, the textual rating description in the assessment grid of the role-play and monologue needs improvement in order to make the marking scale work reliably.

The new description of the criteria was discussed and some phrases in the marking scale were shaped. The Table 11 and Table 12 show the updated speaking test marking scale.*

	Task 1	**Task 2**	**Task 3**	
	Interview	**Role-play**	**Monologue**	
6	Can participate fully in an interview, expanding and developing the point being discussed.	Can use language flexibly and effectively. Can fluently, accurately and spontaneously participate in the given situation.	Can give elaborate narrative, developing particular points and rounding off with an appropriate conclusion.	6
5	Can carry out an effective interview and expand the point being discussed.	Can rather fluently and spontaneously participate in the given situation	Can develop a clear description or narrative, expanding and supporting his/her main points.	5
4	Can develop point of view, but does so with limited precision. (it seems that this description is stricter that which is evaluated for 3 points)	Can maintain a conversation in the given situation but may sometimes be difficult to follow when trying to say exactly what he/she would like to.	Can relate a straightforward narrative stating his/her point of view and comparing.	4
3	Can provide concrete information. (not specified)	Can maintain a conversation in the given situation using simple phrases (using more extended phrases (complex)	Can relate a straightforward narrative. The expression of ideas is simple, sometimes clumsy	3
2	Can provide very simple answers to questions.	Can handle short social exchanges,there are misunderstandings in communication. (seems to be simalar *to misinterprets what is said)*	Can express his/her point of view, but the narrative is clumsy and hard to understand.	2
1	Can reply with very simple answers.	Can handle very short and often inaccurate social exchanges, misinterprets what is said.	Can give separate, very simple, often unrelated statements.	1
0	Not enough to evaluate.			0

Table 35 Revised speaking test marking scales (1)

	Vocabulary	**Grammar**	**Fluency and Pronunciation**	
4	Has a good command of a broad lexical repertoire. Can express him/herself, provide clear descriptions, opinions and arguments.	Consistently maintains a high degree of grammatical accuracy; errors are rare. – not needed	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Has acquired a clear, natural pronunciation and intonation.	4

3	Has a sufficient vocabulary to express him/herself and provide arguments.	Can use various language structures. Occasional „slips" and minor flaws in sentence structure may still occur.	Can express him/herself with ease. Sometimes pauses occur; the pronunciation is clear and easy to understand, errors are rare.	3
2	Has sufficient vocabulary to conduct routine, everyday transactions. (but still systematically lacks words)	Uses reasonably accurately a repertoire of frequently used „routines" . The errors do not cause problems in communication.	Can make him/herself understood in short contributions, even though pauses and reformulation are very evident. Pronunciation is generally clear enough.(even though pauses, reformulation and false starts0	2
1	Can control a narrow lexical repertoire. (narrow and limited lexical repertoire)	Shows only limited control of a few simple grammatical structures and sentence patterns.	Can manage very short utterances, with much pausing to search for expressions, pronunciation can be understood with some effort.	1
0	Not enough to evaluate			0

Table 36 Revised speaking test marking scales (2)

** After the above described analysis, carried out in autumn 2010, the speaking marking scales were piloted again in spring 2011 when the descriptors were discussed and standardized during 9 meetings with over 250 school teachers of English. The final version of the rating scale, used to assess the examination Speaking part in 2011, is available at www. visc.gov.lv.*

Thus, the present research aimed at relating Latvian 2011 Year 12 examination speaking test to the CEFR. The relation was conducted following the guidelines from a manual for relating examinations to the Common European Framework. Therefore, Year 12 examination speaking test was analysed according to the CEFR. Next, the content of the examination was analysed in relation to the relevant categories of the CEFR. For this reason 2011 Year 12 examination speaking test was compared with the CEFR test demands, using the CEFR Grid for speaking developed by ALTE members. Moreover, each task of 2011 Year 12 speaking test was analysed according to the criteria of the CEFR Grid developed by ALTE members. Finally, the specification of content included the examinee response in three tasks in terms of the CEFR. The next stage included benchmarking of a sample speaking test to the level in the CEFR. Moreover, the correspondence of Latvian Year 12 examination level to the CEFR levels was investigated. Finally, Year 12 examination speaking test marking scales were trialled, analysed, compared to the CEFR scales descriptors and consequently some updates to these scales were offered.

References

1. ALTE (2009) ALTE Materials for the Guidance of Test Item Writers. ALTE . Available from www.alte.org

2. Bachman, L. F. (2004) Statistical Analyses for Language Assessment. Cambridge University Press, Cambridge

3. Bachman, L.F. (2001) Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

4. Bachman, L.F., and Palmer, A. (2001) Language testing in practice. Oxford: OUP

5. Bachman, L.F., Davidson, F. Ryan, K., Chol, I. (1995) *Studies in Language Testing 1*: *An investigation into the comparability of two tests of English as a foreign language.* Cambridge: Cambridge University Press.

6. Batstone, R. (1994) Grammar. Oxford: Oxford University Press

7. Betels, Džordžs. Ievads pārbaudes darbu statistiskā analīze. Palīgs skolotājiem

8. Bond, T. and Fox, Ch. (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Hillsdale: Lawrence Erlbaum Associates, Inc.

9. Brown, J. D. (1988). Understanding research in second language learning: A teacher's guide to statistics and research design. London: Cambridge University Press.

10. Brumfit, C. (1984) Communicative Methodology in Language Teaching. Cambridge: Cambridge University Press

11. Buck, G. (2001) Assessing Listening. Cambridge: CUP

12. Council of Europe (2003). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Manual. Preliminary Pilot Version. Language Policy Division, Strasbourg

13. Council of Europe (2009) Relating Language Examinations to the Common European Framework of Reference: Learning, Teaching, Assessment (CEFR). Strasbourg: Language Policy Division

14. Council of Europe, Modern Languages Division (2001) Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

15. Everitt, B.S. (2006) The Cambridge Dictionary of Statistics. Cambridge University Press, Cambridge

16. Figueras N., B. North, S. Takala, N. Verhelst and P. Van Avermaet. Relating examinations to the Common European Framework: a manual, Language testing 2005 22 (3) 261-279, Edward Arnold Publishers, 2005.

17. Fulcher, G. (2003) Testing Second Language Speaking. Malaysia: Pearson Education Limited.

18. Grice, H. P. (1975) Logic and Conversation. In Peter Cole and Jerry L. Morgan (eds.) Syntax and Semantics, Volume 3 (pp. 41-58). New York: Academic Press.

19. Hamp-Lyons, L. (2003) Writing Teachers as Assessors of Writing. In B. Kroll (ed.) Exploring Dynamics of Second Language Writing. (pp. 162-189) Cambridge: Cambridge University Press.

20. Harmer, J. (2001). The Practice of English Language Teaching. Edinburgh: Pearson Education.

21. Hudges, A. (2003) Testing for Language Teachers. Cambridge: Cambridge University Press

22. Hymes, D. (1974) Foundations of Sociolinguistics: An Ethnographic Approach. Philadelphia: University of Pennsylvania.

23. Kaftandjieva, F. (2010) Methods for Setting Cut Scores in Criterion-references Achievement Tests. EALTA

24. Kalnbērziņa, V. (2007) Impact of Relation of Year 12 English Language Examination to CEFR on the Year 12 Writing Test. Proceedings of FIPLV NBR conference.

25. Lado, R. (1965) Language Testing – The Construction and Use of Foreign Language Tests. London: Lowe and Brydone. 4th ed.

26. Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? [Online]. Rasch Measurement Transactions, 16(2): 878. Available from http://www.rasch.org/rmt/rmt162f.htm

27. Lumley, T., & Brown, A. (2005). Research Methods in Language Testing. In Hinkel, E. (ed.) Handbook of Research in Second Language Teaching and Learning. (pp. 833-836) Hillsdale, N.J.: Lawrence Erlbaum Associates

28. Luoma, S. (2004). Assessing speaking. Cambridge: Cambridge University Press.

29. McNamara, T. (1996). Measuring Second Language Performance. Harlow: Addison Wesley Longman Ltd.

30. Mohamad, A. (1999) What Do We Test When We Test Reading Comprehension? The Internet TESL Journal, V (12) Available at http://iteslj.org/Techniques/Mohamad-TestingReading.html [Accessed January 23, 2011].

31. Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

32. Nunan, D. (2004) Research Methods in Language Learning. Cambridge: Cambridge University Press.

33. Nuttall, Ch. (2005) Teaching Reading Skills in a Foreign Language. Oxford: Macmillan.

34. Oermann, M., Gaberson, K. (1998) Evaluation and testing in nursing education. Springer Publishing Company, Inc.

35. Rasinger, S. M. (2010) Quantitative Methods: Concepts, Frameworks and Issues. In L. Litosseliti (ed.) Research Methods in Linguistics. (pp. 49-67) London: Continuum.

36. Richards, J.C. and Rodgers, T. S. (2001) Approaches and Methods in Language Teaching. Cambridge: Cambridge University Press.

37. Tamrackitkun, K. (2010) Extensive Reading: An Empirical Study of Its Effects on EFL Thai Students' Reading Comprehension, Reading Fluency and Attitudes. Doctoral Thesis. Salford: University of Salford.

38. Thornbury, S. (2002). How to Teach Vocabulary. Harlow: Longman.

39. Weigle, S. C. (2002) Assessing Writing. Cambridge: Cambridge University Press.

40. Widdowson, H.G. (1978) Teaching Language as Communication. Oxford: Oxford University Press

41. Wright, B., & Stone., M. (1999) Measurement Essentials [Online]. Wilmington: Wide Range. Available from http://www.rasch.org/memos.htm#measess. Accessed May 2, 2008

Internet sources:

1. British Council. Available from: http://www.britishcouncil.org/ielts-coe_flier.pdf [Accessed January 4, 2011].

2. CEFR standards. Retrieved from http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html and

3. http://visc.gov.lv/eksameni/vispizgl/uzdevumi/2010/vidussk/12kl_angl_val.pdf
4. http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20INput51.pdf[Accessed January 4, 2011].

5. http://www.scribd.com/doc/47074250/Iteman-Manual
6. http://www.visc.gov.lv/eksameni/vispizgl/dokumenti/20101029_par_ce_paraugiem_svesvalodas.pdf [Accessed January 4, 2011].

7. ISEC (2009) *Centralizētā eksāmena angļu valodā rakstīšanas daļas vērtēšana: Metodiskais materiāls*. Rīga: Izglītības satura un eksaminācijas centrs. Available from www.isec.gov.lv/eksameni/vispizgl/.../angl_valoda_rakst_met_mat.pdf [Accessed January 23, 2011].
8. ISEC (2010) *Par centralizēto eksāmenu paraugiem svešvalodās (angļu, franču, krievu un vācu valodā).* Rīga: Izglītības satura un eksaminācijas centrs. Available from www.isec.gov.lv/eksameni/.../20101029_par_ce_paraugiem_svesvalodas.pdf [Accessed January 23, 2011].
9. O'Sullivan, B. (2008) *Notes on Assessing Speaking.* Available from: http://www.lrc.cornell.edu/events/past/2008-2009/papers08/osull1.pdf [Accessed January 4, 2011].

10. Vispārējās vidējās izglītības mācību priekšmeta standarts Available from: http://www.likumi.lv/doc.php?id=181216 [Accessed January 4, 2011].

Appendices

Appendix 1 Year 12 Examination in English 2010 Reading Part Fit statistics

TABLE 10.1 ANGL.xls
INPUT: 22637 PERSON 30 ITEM MEASURED: 22637 PERSON 30 ITEM 60 CATS WINSTEPS 3.70.0.5
PERSON: REAL SEP.: 2.40 REL.: .85 ... ITEM: REAL SEP.: 48.56 REL.: 1.00

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBISERL-EX CORR.	PTBISERL-EX EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ESTIM DISCR	ITEM	G
19	7277	22637	.67	.02	1.46	9.9	1.81	9.9	A .05	.43	65.7	76.2	.19	209	0
2	4637	22637	1.47	.02	1.34	9.9	1.58	9.9	B .14	.40	76.5	82.6	.61	102	0
5	3569	22637	1.88	.02	1.21	9.9	1.57	9.9	C .19	.38	84.3	85.8	.77	105	0
18	13138	22637	-.73	.02	1.34	9.9	1.50	9.9	D .12	.40	56.2	70.2	.10	208	0
13	9136	22637	.20	.02	1.32	9.9	1.45	9.9	E .18	.43	62.3	72.9	.33	203	0
15	7732	22637	.55	.02	1.32	9.9	1.41	9.9	F .19	.43	65.0	75.3	.46	205	0
16	8089	22637	.46	.02	1.21	9.9	1.33	9.9	G .26	.43	68.4	74.6	.59	206	0
12	10070	22637	-.02	.02	1.16	9.9	1.23	9.9	H .30	.43	66.3	71.6	.63	202	0
20	14179	22637	-.97	.02	1.15	9.9	1.20	9.9	I .26	.38	64.7	70.9	.64	210	0
11	10929	22637	-.22	.02	1.12	9.9	1.17	9.9	J .32	.42	66.2	70.8	.70	201	0
14	18498	22637	-2.15	.02	.98	-2.2	.95	-2.0	K .30	.28	82.4	82.3	1.03	204	0
24	9037	22637	.23	.02	.94	-8.2	.91	-8.5	L .48	.43	74.9	73.0	1.12	304	0
26	9292	22637	.16	.02	.94	-9.1	.91	-9.3	M .48	.43	74.7	72.6	1.13	306	0
1	12872	22637	-.66	.02	.93	-9.9	.88	-9.9	N .46	.40	72.6	70.1	1.19	101	0
22	7315	22637	.66	.02	.92	-9.9	.89	-9.2	O .49	.43	78.2	76.1	1.13	302	0
9	12964	22637	-.69	.02	.91	-9.9	.86	-9.9	o .48	.40	74.3	70.1	1.25	109	0
17	17030	22637	-1.70	.02	.89	-9.9	.76	-9.9	n .41	.32	79.0	77.0	1.19	207	0
30	9619	22637	.09	.02	.89	-9.9	.85	-9.9	m .52	.43	76.4	72.2	1.24	310	0
6	8441	22637	.37	.02	.88	-9.9	.85	-9.9	l .53	.43	77.8	74.0	1.22	106	0
27	10633	22637	-.15	.02	.86	-9.9	.80	-9.9	k .54	.42	76.0	71.0	1.33	307	0
7	9811	22637	.04	.02	.86	-9.9	.82	-9.9	j .55	.43	77.1	71.9	1.31	107	0
28	9026	22637	.23	.02	.86	-9.9	.82	-9.9	i .55	.43	78.2	73.0	1.28	308	0
25	7625	22637	.58	.02	.85	-9.9	.82	-9.9	h .54	.43	80.9	75.5	1.24	305	0
8	12072	22637	-.48	.02	.84	-9.9	.77	-9.9	g .55	.41	77.2	70.1	1.43	108	0
10	12240	22637	-.52	.02	.84	-9.9	.77	-9.9	f .55	.41	76.9	70.1	1.44	110	0
3	12178	22637	-.51	.02	.83	-9.9	.76	-9.9	e .55	.41	77.2	70.1	1.45	103	0
23	6439	22637	.90	.02	.83	-9.9	.76	-9.9	d .56	.43	82.4	78.0	1.25	303	0
29	7205	22637	.69	.02	.81	-9.9	.78	-9.9	c .58	.43	82.8	76.3	1.30	309	0
4	10015	22637	-.01	.02	.80	-9.9	.74	-9.9	b .59	.43	79.4	71.7	1.45	104	0
21	11708	22637	-.40	.02	.79	-9.9	.72	-9.9	a .59	.41	79.7	70.3	1.54	301	0
MEAN	10092.5	22637	.00	.02	1.00	-3.0	1.02	-3.0			74.4	73.9			
S.D.	3211.3	.0	.81	.00	.20	9.2	.31	9.2			6.8	4.0			

Appendix 2 Year 12 Examination in English 2010 Language Use Part Fit Statistics

TABLE 10.1 ANGV.xls

INPUT: 22637 PERSON 45 ITEM MEASURED: 22637 PERSON 45 ITEM 90 CATS WINSTEPS 3.70.0.5

PERSON: REAL SEP.: 2.68 REL.: .88 ... ITEM: REAL SEP.: 67.72 REL.: 1.00

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBISERL-EX CORR.	PTBISERL-EX EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM	G
27	17083	22637	-1.54	.02	1.15	9.9	1.62	9.9	A .13	.30	73.7	76.4	209	0
18	6067	22637	1.05	.02	1.21	9.9	1.45	9.9	B .18	.39	75.0	77.8	118	0
24	10766	22637	-.08	.01	1.30	9.9	1.41	9.9	C .11	.39	55.4	68.6	206	0
9	5785	22637	1.13	.02	1.15	9.9	1.41	9.9	D .22	.39	77.4	78.6	109	0
21	14690	22637	-.95	.02	1.09	9.9	1.32	9.9	E .24	.34	67.9	70.0	203	0
16	16783	22637	-1.46	.02	1.09	9.9	1.30	9.9	F .19	.30	75.0	75.4	116	0
14	8758	22637	.37	.02	1.18	9.9	1.28	9.9	G .22	.40	66.1	71.4	114	0
3	5746	22637	1.14	.02	1.04	4.2	1.25	9.9	H .32	.39	80.0	78.8	103	0
2	13440	22637	-.66	.01	1.05	9.0	1.24	9.9	I .29	.36	68.6	68.4	102	0
1	14900	22637	-.99	.02	1.15	9.9	1.23	9.9	J .20	.34	64.0	70.4	101	0
23	10613	22637	-.05	.01	1.14	9.9	1.20	9.9	K .25	.39	63.0	68.7	205	0
45	13064	22637	-.58	.01	1.02	4.6	1.14	9.9	L .32	.36	69.0	68.1	317	0
12	8426	22637	.45	.02	1.07	9.6	1.12	9.9	M .33	.40	71.0	72.1	112	0
4	10358	22637	.01	.01	1.09	9.9	1.11	9.9	N .31	.39	65.8	69.0	104	0
38	3509	22637	1.91	.02	1.08	6.2	.92	-3.7	O .34	.36	84.1	86.0	310	0
15	12480	22637	-.45	.01	1.06	9.9	1.07	7.1	P .31	.37	64.4	67.9	115	0
10	7883	22637	.58	.02	1.04	5.4	1.04	4.2	Q .36	.40	71.9	73.2	110	0
39	4945	22637	1.39	.02	1.03	2.8	1.02	1.1	R .35	.38	81.1	81.2	311	0
25	15573	22637	-1.15	.02	1.02	3.9	1.02	1.2	S .31	.33	70.7	71.9	207	0
20	18712	22637	-2.04	.02	.99	-1.1	1.02	.8	T .26	.26	83.0	82.9	202	0
26	17225	22637	-1.58	.02	1.01	1.4	.98	-.9	U .28	.29	77.2	76.9	208	0
31	4253	22637	1.62	.02	1.00	-.1	.92	-4.3	V .38	.37	83.2	83.5	303	0
17	11734	22637	-.29	.01	.98	-3.9	.98	-1.8	W .40	.38	69.3	68.0	117	0
8	5609	22637	1.18	.02	.98	-2.0	.96	-3.0	v .41	.39	78.8	79.2	108	0

	36	7142	22637	.76	.02	.98	-2.5	.95	-4.1	u	.42	.40	75.1	74.9	308	0
	11	10654	22637	-.06	.01	.97	-5.7	.95	-6.0	t	.42	.39	70.3	68.7	111	0
	28	17877	22637	-1.77	.02	.97	-3.5	.92	-4.2	s	.31	.28	79.9	79.4	210	0
	19	16666	22637	-1.43	.02	.97	-4.4	.89	-6.6	r	.34	.31	75.7	75.0	201	0
	7	10636	22637	-.05	.01	.96	-6.9	.95	-5.7	q	.42	.39	70.8	68.7	107	0
	22	19699	22637	-2.42	.02	.96	-3.2	.84	-5.9	p	.27	.23	87.2	87.1	204	0
	5	13098	22637	-.59	.01	.93	-9.9	.93	-6.9	o	.42	.36	72.0	68.1	105	0
	42	9170	22637	.27	.02	.93	-9.9	.90	-9.9	n	.47	.40	72.1	70.7	314	0
	29	4172	22637	1.65	.02	.93	-6.5	.86	-8.0	m	.43	.37	84.9	83.7	301	0
	35	4209	22637	1.64	.02	.92	-6.9	.76	-9.9	l	.46	.37	84.2	83.6	307	0
	32	4722	22637	1.46	.02	.91	-8.6	.76	-9.9	k	.48	.38	82.5	81.9	304	0
	13	11284	22637	-.19	.01	.91	-9.9	.87	-9.9	j	.47	.38	72.4	68.2	113	0
	33	15673	22637	-1.18	.02	.88	-9.9	.81	-9.9	i	.43	.32	76.5	72.1	305	0
	34	8374	22637	.46	.02	.88	-9.9	.84	-9.9	h	.51	.40	76.2	72.2	306	0
	43	7618	22637	.64	.02	.86	-9.9	.80	-9.9	g	.53	.40	78.1	73.8	315	0
	30	14924	22637	-1.00	.02	.84	-9.9	.79	-9.9	f	.48	.34	78.1	70.4	302	0
	40	5636	22637	1.17	.02	.84	-9.9	.71	-9.9	e	.55	.39	81.9	79.1	312	0
	41	3215	22637	2.03	.02	.82	-9.9	.64	-9.9	d	.50	.35	89.1	87.1	313	0
	6	12387	22637	-.43	.01	.82	-9.9	.78	-9.9	c	.54	.37	78.2	67.9	106	0
	37	8652	22637	.39	.02	.81	-9.9	.75	-9.9	b	.58	.40	78.7	71.6	309	0
	44	12066	22637	-.36	.01	.78	-9.9	.72	-9.9	a	.59	.37	79.4	67.9	316	0
MEAN		10583.9	22637	.00	.02	1.00	-.4	1.01	-.8				75.1	74.6		
S.D.		4608.6	.0	1.14	.00	.12	8.0	.22	8.2				7.0	6.0		

Appendix 3 Year 12 Examination in English 2010 Listening Part Fit Statistics

INPUT: 22637 PERSON 30 ITEM MEASURED: 22637 PERSON 30 ITEM 60 CATS WINSTEPS 3.70.0.5

PERSON: REAL SEP.: 2.14 REL.: .82 ... ITEM: REAL SEP.: 68.34 REL.: 1.00

ITEM STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBISERL-EX CORR.	PTBISERL-EX EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM	G
22	5140	22637	1.84	.02	1.15	9.9	1.54	9.9	A .16	.33	78.6	79.6	302	0
13	3336	22637	2.47	.02	1.06	4.9	1.47	9.9	B .20	.29	85.8	85.9	203	0
24	5079	22637	1.86	.02	1.10	9.9	1.45	9.9	C .20	.33	79.5	79.8	304	0
25	13234	22637	-.12	.02	1.20	9.9	1.37	9.9	D .18	.37	62.5	70.1	305	0
11	21577	22637	-3.20	.03	1.01	.6	1.36	6.7	E .13	.17	95.4	95.4	201	0
29	5266	22637	1.80	.02	1.03	3.6	1.36	9.9	F .25	.34	80.4	79.2	309	0
15	12695	22637	.00	.02	1.22	9.9	1.31	9.9	G .18	.38	60.7	69.7	205	0
19	17365	22637	-1.16	.02	1.12	9.9	1.29	9.9	H .19	.32	76.5	78.2	209	0
27	16437	22637	-.90	.02	1.13	9.9	1.24	9.9	I .21	.34	72.6	75.5	307	0
30	9312	22637	.76	.02	1.15	9.9	1.22	9.9	J .25	.38	64.9	70.9	310	0
28	11640	22637	.24	.01	1.16	9.9	1.21	9.9	K .24	.38	62.7	69.5	308	0
23	13242	22637	-.12	.02	1.12	9.9	1.17	9.9	L .26	.37	64.9	70.1	303	0
20	15012	22637	-.54	.02	1.08	9.9	1.17	9.9	M .27	.36	70.4	72.3	210	0
12	13466	22637	-.18	.02	1.10	9.9	1.08	8.0	N .29	.37	64.7	70.2	202	0
17	8908	22637	.85	.02	1.04	5.9	1.05	5.4	O .34	.38	70.1	71.3	207	0
26	11553	22637	.26	.01	1.03	5.4	1.03	3.3	o .35	.38	67.9	69.5	306	0
16	16419	22637	-.90	.02	1.02	2.8	1.01	.4	n .32	.34	74.8	75.4	206	0
18	15034	22637	-.54	.02	.99	-1.1	.91	-7.9	m .38	.36	70.5	72.3	208	0
14	18744	22637	-1.61	.02	.97	-3.1	.84	-7.8	l .33	.29	83.4	83.3	204	0
2	7856	22637	1.10	.02	.90	-9.9	.82	-9.9	k .47	.37	75.6	72.9	102	0
8	12130	22637	.13	.01	.89	-9.9	.86	-9.9	j .48	.38	74.1	69.6	108	0
3	9750	22637	.66	.02	.86	-9.9	.81	-9.9	i .51	.38	75.6	70.4	103	0
5	17098	22637	-1.08	.02	.85	-9.9	.76	-9.9	h .46	.32	80.9	77.4	105	0

```
|     9   9081  22637     .81     .02| .85  -9.9| .79  -9.9|g .53   .38| 76.6  71.1| 109   0 |
|     4  19161  22637   -1.76     .02| .85  -9.9| .66  -9.9|f .42   .27| 85.6  84.9| 104   0 |
|    21  13819  22637    -.26     .02| .83  -9.9| .76  -9.9|e .53   .37| 77.3  70.6| 301   0 |
|     1  12087  22637     .14     .01| .79  -9.9| .73  -9.9|d .58   .38| 78.9  69.6| 101   0 |
|     6  17836  22637   -1.30     .02| .79  -9.9| .62  -9.9|c .51   .31| 82.6  79.8| 106   0 |
|    10  10299  22637     .54     .02| .78  -9.9| .72  -9.9|b .58   .38| 79.2  70.0| 110   0 |
|     7  11737  22637     .21     .01| .76  -9.9| .70  -9.9|a .61   .38| 80.3  69.5| 107   0 |
|-----------------------------------+----------+----------+-----------+-----------+---------|
| MEAN 12477.1  22637     .00     .02| .99    .6|1.04    .6|           | 75.1  74.8|         |
| S.D.  4484.1     .0    1.19     .00| .14   8.7| .27   9.2|           |  7.9   6.3|         |
-------------------------------------------------------------------------------------------------
```

List of Tables