DE GRUYTER OPEN

10.1515/rjes -2016-0015

SYNTACTIC AND LEXICAL COMPLEXITY OF B2 LISTENING COMPREHENSION SUBTESTS IN ENGLISH: A COMPARATIVE STUDY

GAŠPER ILC
University of Ljubljana, Slovenia
ANDREJ STOPAR
University of Ljubljana, Slovenia

Abstract: *Adopting Weir's (2005) socio-cognitive validation framework, the present paper focuses on the syntactic and lexical complexity of listening comprehension subtests in three B2-level examinations: The City Guilds international examination in English, The First Certificate in English, and the General Matura in English. By analysing and interpreting the results obtained from different automated tools, the research aims to determine to what extent the three subtests are comparable. The results of the study suggest the unreliability of the Common European Framework of Reference for Languages (CEFR) as a sole mechanism for test comparisons.*
Keywords: *context validity, lexical complexity, listening comprehension texts, syntactic complexity*

1. Introduction

The question of test comparability has long been topical in the field of language testing. Since its publication, the Common European Framework of Reference (CEFR) has been increasingly used to facilitate attempts to compare the target levels of different language tests. Ostensibly, by establishing common levels of reference, the CEFR made it possible for both the test-developers and test-takers to identify the proficiency levels expected, tested and awarded. However, since tests are designed to meet different objectives and to assess knowledge of different populations, it remains questionable to what extent two (or more) tests targeting the same CEFR level are truly comparable.

As it is an established fact that any comparability analysis of language tests should encompass a wide array of factors (Taylor 2004), a number of proposals that also entail components not necessarily included in the CEFR have been developed. An oft-quoted example of such a model is Weir's (2005) socio-cognitive validation framework, which focuses on different types of test score validity. The aim of the analysis presented herein is to focus only on one segment of this elaborate validation framework in three B2-level examinations: the City Guilds international examination in English (C&G), The First Certificate in English (FCE), and the General Matura in English (GM). Our main objective is to establish to what extent the three examination subtests are comparable in terms of their syntactic and lexical complexity and to investigate how these factors relate to the test-providers' claims that their test targets the B2 CEFR level. The analysis mostly draws on the results obtained from automated tools (VocabProfile 6.2, Coh-Metrix Version 3.0).

2. Test (Score) Validity: Basic Tenets

In the field of language assessment, the concept of test validity pertains to the critical exploration of how meaningful a test score is. According to Weir (2005:12), the most basic definition of validity, which continues to be relevant today, can be traced to psychologist Truman L. Kelley, who already in 1927 stated that "the problem of validity is that of whether a test really measures what it purports to measure" (quoted in Weir 2005:12). Although some researchers and practitioners in the field still subscribe to this basic, unitary description of validity, there are those who have elaborated the concept further and developed models that distinguish between several types or aspects of validity (e.g., American Psychological Association et al. 1954; Messick 1995; Weir 2005).

As a non-unitary approach, Weir's (2005) socio-cognitive framework enables a broad comparison of heterogeneous tests, and thus takes into account the oft-reported inadequacy of relying solely on CEFR descriptors when comparing tests that are based on different constructs (see Alderson et al. 2004; Jones 2009; Ilc and Stopar 2015 among others). This framework draws on evidence from different types of validity, five of which are presented below.

Perhaps the most fundamental type of validity is "theory-based validity" (also referred to as "construct validity") as it is related to ensuring a match between the theoretical construct and the test. For instance, if the listening comprehension construct includes only listening for detail, the test should not elicit answers measuring other listening skills, such as listening for main idea. Related to this concept is "criterion-related validity", which is established by correlating test performances with a valid external measure of the same construct – usually this can be some other test measuring the same skill. To continue, "scoring validity" covers what some authors describe as "reliability", i.e., it refers to statistical analyses of the test that provide the test providers information with regard to the degree to which they depend on the test results. Another type of validity is "consequential validity", which takes into account the wider social consequences of testing: the perception of the test scores by the stakeholders, the backwash of the test, and such. See Weir (2005) for more on these types of validity.

For the purposes of the analysis presented in this article, it is assumed that the listening subtests under investigation here are not disputable with regard to the types of validity described thus far. Namely, the C&G, the FCE, and the GM are analysed merely from the perspective of "context validity", which is sometimes also referred to as "content validity". Weir (2005:19) defines it as follows:

> Context validity is concerned with the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting.

Broadly speaking, a contextually valid test score is feasible if the test situation is made as similar as possible to real-life situations. In line with Weir's definition above, this includes taking into account a variety of linguistic, content-related and other demands, such as: the rubric; authenticity of texts and activities; selection of task-types; ordering of items in a task; assessment information; the setting for and the administration of the test; and various features of the texts selected for the examination such as length, genre, structure, lexical difficulty, structural demands, etc. Some other contextual factors also include the constraints on timing, speech rates, the nature of information in the text (abstract vs. concrete), the amount of background/content knowledge required to do the tasks, the type of input provided, the type of output required, the communicative functions required, and so forth (see Weir 2005 for

additional examples). The emphasis of the present analysis is, however, predominantly on lexical and structural components of context validity.

3. Contextually Valid Texts: Lexical and Syntactic Aspects

With regard to the selection of contextually valid listening comprehension texts, text features such as vocabulary selection, text/task length, and syntactic complexity are paramount.

Buck (2001:15) defines the process of understanding words as including the basic stages of recognising the word and understanding its meaning. As this process also relies on contextual knowledge, word-frequency is a significant factor to consider in the selection of texts. Buck (2001:16) summarizes Garnham's (1985) findings that "higher-frequency words are recognised faster than lower frequency words," and that "word recognition is faster if the words are in a helpful context." Related to this, Buck (2001:170) also warns that slang and less frequent words should be avoided, especially when testing lower-ability students. Weir (2005:78) further substantiates these claims by quoting Bond and Garnes, who showed that "[i]n listening, low-frequency lexical items are less likely to be recognized or more likely to be misheard." Consequently, Weir (2005:78) suggests that item writers should always investigate whether "the input or out text require knowledge of too many unknown lexical items." While recent research (e.g., Matthews and Cheng 2015) suggests that high-frequency words only partially account for the variance in the listening comprehension scores, the negative impact of low-frequency/unknown vocabulary seems to remain undisputed. Kobeleva's (2012) study, for instance, shows that even the presence of unknown proper names in a listening comprehension text can negatively affect the performance of test-takers.

The length of texts is frequently perceived as an important element of listening comprehension tests (e.g., Alderson 2006 and Rost 2006). However, according to Bloomfield et al. (2011:2317-2318), empirical studies in second language listening comprehension so far have failed to show a strong connection between passage length and item difficulty. The reason for this lies in the fact that text length is not always indicative of the amount of information a text contains. In addition, other factors, such as the amount of repetition, discourse type and speech rate, make it difficult to make any conclusive claims about the effect of text length alone. In spite of this, research has shown that "[l]onger texts will tend to require discourse skills, whereas shorter texts will tend to focus more on localised grammatical characteristics" (Buck 2001:123). Thus, the length of the text in most cases determines the type of skills and strategies you can test (Weir 2005:74).

Structural characteristics are also a contributing factor when tackling context validity. In his seminal book on assessing reading, Alderson (2000:37) emphasizes the importance of "the ability to parse sentences into their correct syntactic structure," while Shiotsu and Weir (2007) even produce evidence that syntactic knowledge may be more important for predicting reading test performance than vocabulary. Similar claims have been made for the listening skill: already in 1990, Rost pointed out the role of listener familiarity with syntactic patterns used in the listening text and concluded that unexpected patterns may hinder comprehension. Thus, when considering listening comprehension, claims Buck (2001:10), it should be taken into account that in spoken language idea units tend to be shorter with simpler syntax; for instance, they typically contain fewer dependent and subordinate clauses. Also important in this respect are various connectives – their role as discourse markers has been widely discussed since the ground-breaking work of Halliday and Hasan (1976) on cohesion in English. Buck (2001:10) states that, in listening texts, idea units in spoken language are strung together relatively simply by coordinating conjunctions such as "and", "or", and "but". Furthermore, McNamara et al. (2014) observe that the overall high frequency of connectives

and discourse markers adds positively to the cohesion and text ease. In particular, the authors (2014:36) point out that this is especially the case with the causal connectives (e.g., "because" and "so") and logical connectives (e.g., "or", "if" and "then"): high incidence of causal and logical connectives is directly linked to high cohesion.

To conclude the introductory section, let us quote Alderson's observation (2000:70-71) on some aspects of contextual validity: "[c]learly at some level the syntax and lexis of texts will contribute to text and thus difficulty, but the interaction among syntactic, lexical, discourse and topic variables is such that no one variable can be shown to be paramount." We believe that our validation study adds positively to a better understanding of these factors and their interdependencies.

4. Study

4.1. The Context

The present study examines three listening comprehension subtests, for which the test providers claim that they are aligned with the B2 level of the CEFR scale. The subtests under investigation are: (i) The First Certificate in English (FCE), (ii) The City Guilds international examination in English (C&G), and (iii) the General Matura in English (GM). The first two are ESOL examinations administered internationally, whereas the GM is a secondary school leaving examination administered on the national level. The structure of the three subtests is presented in Table 1.

		Task-Type	**Text-Type**	**Items**	**Target**
FCE	Part 1	multiple-choice	8 short dialogues	8	gist/detail
	Part 2	gap-fill	monologue	10	detail
	Part 3	matching	5 short monologues	5	gist
	Part 4	multiple-choice	interview	7	overall comp. ability
C&G	Part 1	multiple-choice	8 short dialogues	8	gist
	Part 2	multiple-choice	3 dialogues	6	gist
	Part 3	completion	monologue	8	detail
	Part 4	multiple-choice	conversation	8	overall comp. ability
GM	Part 1	T/F	interview	9	detail
	Part 2	short answers	interview	8	overall comp. ability

Table 1. The internal structure of the listening subtests

4.2. Instruments

To determine the syntactic and lexical complexity of the three listening comprehension subtests, two automated text-analysis tools were used: VocabProfile version 6.2 (Cobb, n.d.; Heatley et al. 2002), and Coh-Metrix Version 3.0 (McNamara et al. 2005). VocabProfile version 6.2 was applied to determine the lexical frequency and consequently lexical complexity, whereas Coh-Metrix Version 3.0 was employed to detect lexical diversity, word information, connectives, syntactic complexity, syntactic pattern density, and readability. All of the listed categories play an important role in determining the context validity within Weir's (2005) validation framework.

Prior to the analysis, the three listening subtests were word-processed and saved as plain text. Since they involve reading comprehension as well (i.e., understanding questions),

each subtest was divided into two parts: (i) the questions (Q), and (ii) the tapescripts (T) to minimize the influence of the reading input on the final results.

4.3. Results

VocabProfile version 6.2 was applied to the three subtests for questions and tapescripts. The results of the analysis are presented in Table 2.

	FCE		C&G		GM	
BNC levels	**Q**	**T**	**Q**	**T**	**Q**	**T**
k1	83.13	87.02	89.00	91.91	68.15	84.01
k2	8.35	5.90	4.63	3.39	7.01	5.53
k3	1.70	1.87	2.12	1.18	6.37	2.36
k4	0.51	0.73	0.39	0.28	3.82	0.94
k5	-	0.24	0.58	0.45	-	0.67
k6	0.68	0.28	-	0.11	0.64	0.34
k7	0.68	0.85	0.19	-	0.64	0.34
k8	1.36	0.57	0.39	0.06	-	0.61
k9	-	-	-	0.11	-	0.13
k10	-	-	-	0.06	-	0.07
k11	-	-	-	-	-	0.27
k12	-	-	-	-	-	-
k13	-	-	-	-	-	0.07
k14	-	-	-	-	-	0.07
k15-k20	-	-	-	-	-	-
OFF-list	3.58	2.52	2.51	1.91	13.38	4.39

Table 2. Values of the VocabProfile analysis

The k1– k20 levels in Table 2 correspond to the word list frequency levels as set by the British National Corpus: level k1 (1,000-word level) contains basic and frequently used words, whereas level k20 (20,000-word level) contains low frequency words. The Vocabprofile analysis shows that the tapescripts of the three listening subtests mostly contain k1, k2 and k3 lexical items (from 92%-96% overall). Proportionally, the GM has a slightly lower percentage of the k1 items, and the highest percentage of the k3 items. The GM also contains some lexical items belonging to categories k11, k13 and k14, as well as several geographical and proper names which the Vocabprofile listed as off-list (e.g., *Zanzibar, Shirley*). The C&G has the highest number of k1 items (91.91% in the T section); however, contrary to the FCE, which is limited to k1-k8 items, the G&C also contains k9 and k10 items. The vocabulary of the Q section is comparable to that of the T section, and in none of the analysed subtests are the lexical items more difficult in the Q section than in the T section.

Next, the Coh-Metrix analysis of the three subtests (questions and tapescripts separately) was applied, the results of which are presented in Table 3.

	FCE	C&G	GM

	Q	T	Q	T	Q	T
DESCRIPTIVE						
Word count, number of words	588	2461	528	1795	160	1488
LEXICAL DIVERSITY						
Lexical diversity, type-token ratio, all words	0.462	0.322	0.471	0.317	0.648	0.346
Lexical diversity, MTLD, all words	62.835	110.98	53.619	90.714	77.345	66.924
Lexical diversity, VOCD, all words	83.966	128.99	106.806	116.314	89.15	89.8
CONNECTIVES, incidence						
All connectives	49.32	87.769	49.242	88.579	12.50	94.086
Causal connectives	28.912	33.32	20.833	26.741	25.00	18.817
Logical connectives	20.408	44.697	28.409	52.925	6.25	35.618
Adversative and contrastive connectives	1.701	14.222	9.47	21.17	6.25	14.113
Temporal connectives	11.905	18.285	11.364	25.07	0,00	14.785
Expanded temporal connectives	25.51	26.818	26.515	25.07	12.50	14.785
Additive connectives	10.204	38.196	22.727	35.655	0	65.86
SYNTACTIC COMPLEXITY						
Left embeddedness, words before main verb, mean	1.176	0.93	0.702	1.564	2.647	2.273
Number of modifiers per noun phrase, mean	0.71	0.516	0.694	0.474	1.116	0.671
SYNTACTIC PATTERN DENSITY, incidence						
Noun phrase density	365.646	386.835	371.212	372.145	350	387.769
Verb phrase density	246.599	239.334	253.788	223.955	156.25	178.091
Adverbial phrase density	11.905	39.821	51.136	69.638	6.25	50.403
Preposition phrase density	129.252	92.239	49.242	75.766	100.00	117.608
Agentless passive voice density	11.905	2.032	1.894	0.557	6.25	4.032
Negation density	5.102	8.939	7.576	9.471	0.00	9.409
WORD INFORMATION, incidence						
Noun	250	223.487	276.515	191.087	368.75	229.166
Verb	153.061	134.905	181.818	134.819	131.25	118.279
Adjective	54.421	67.453	66.288	62.953	93.75	55.107
Adverb	25.511	61.356	66.288	105.849	6.25	76.613
READABILITY						
Flesch Reading Ease	74.182	83.766	85.15	92.149	51.854	80.149
Flesch-Kincaid Grade level	4.631	3.288	2.467	2.242	8.365	4.968
Coh-Metrix L2 Readability	24.92	15.093	29.60	16.795	23.527	15.514

Table 3. Values of the Coh-Metrix analysis

As table 3 shows, most values of the Coh-Metrix analysis are comparable; however, some differences can be observed in individual categories. To start with the length of the tapescripts (columns T), the GM is the shortest and the FCE the longest subtest. In contrast, if we calculate the ratio between the length of the tapescripts and the number of tasks (see Table 1), the situation changes: the GM has approximately 744 words per task, the FCE 615 words per task, and the C&G 448 words per task.

With regard to lexical density, in the subcategory of the type-token ratio, the GM has the highest score, but its values are the lowest in the lexical diversity indices of MTLD (measure of textual lexical diversity) and VOCD (vocabulary diversity). The FCE has the highest values in both these categories, while the C&G displays median scores with regard to lexical diversity.

The GM has the highest incidence in the "all connectives" category, and this relatively high score can be mostly attributed to the high incidence of the simple, additive connectives (e.g., *and, also, moreover*). In the category of the causal and logical connectives, the GM has the lowest score. The highest incidence of the causal connectives can be observed in the FCE, whereas the C&G has the highest incidence score with logical, adversative/contrastive and temporal connectives.

With regard to left-embeddedness, extremely high values can be observed in both categories for the GM subtest. The values of the number of modifiers per noun phrase are more equally distributed in the T sections, but there is again a noticeable difference in the Q section, since the GM again has the highest value of the three.

In terms of syntactic pattern density, the FCE displays average values in most subcategories, whereas the values of the C&G and the GM oscillate from the highest to the lowest. The GM has the highest noun phrase density incidence but at the same time the lowest verb phrase density incidence in the T section. There appears to be a direct correspondence between phrase density incidence and word incidence: the subtest with the highest noun incidence value also has the highest noun phrase density value and vice versa. Agentless passive voice density is the highest in the FCE subtest, and the lowest in the C&G, which in turn has the highest negation incidence value.

In the readability category, the Flesch Reading Ease shows that all three subtests display a higher value in the Q section than in the T section (a high value indicates an easier text, and a low value a more difficult text.). However, while the values of the Q and the T sections of the FCE and the C&G subtests are close together, there is a considerable difference in the case of the GM. The values of the Coh-Metrix L2 Readability are very similar in all three subtests, and in all three cases the values in the Q sections are higher than those of the T section.

5. Discussion

The lexical analysis (Vocabprofile) shows that the GM listening subtest is a relative outlier among the three B2 listening subtests. It has the lowest percentage of the k1 items, and the highest percentage of the k3 items; also, it contains lexical items belonging to categories k11, k13 and k14, as well as several geographical and proper names listed as off-list items (specialised vocabulary not present in the BNC classification). The C&G and the FCE, on the other hand, are similar in this respect. Since frequency and familiarity of the lexis are significant factors affecting the difficulty of the test (see Buck 2001; Kobeleva 2012; and Matthews and Cheng 2015), the results can be interpreted as showing that the GM is lexically the most demanding of the three tests analysed.

Turning to the Coh-Metrix analysis, we can observe some differences in individual categories, even though that in most categories the results of the three subtests are comparable.

With regard to text length, the GM is the shortest and the FCE the longest subtest. However, a closer look at the tapescripts (words/tasks ratio, see section 4 above) reveals that GM test-takers are exposed to two longer texts, whereas the FCE and the C&G contain a series of shorter texts/tasks. The question arises whether this is done intentionally – in other words, do the task writers choose shorter/longer texts in order to measure different discourse

skills and listening strategies (see Buck 2001:123). The issue is especially relevant for the GM since its listening subtest relies on merely two relatively long texts to measure either comprehension of details or overall comprehension. Such an approach to listening comprehension is problematic from more than one point of view. Firstly, there may not be enough focus on the skills and/or strategies that the subtest aims to measure; and secondly, the lack of variety that it entails can exacerbate the problems of the test-takers not familiar with the topic, text-type or task-type. Furthermore, a relatively low number of items in the GM subtest (17 items versus 30 items in the C&G and the FCE subtests, see Table 1) can only add negatively to the overall reliability of the listening subtest.

According to McNamara et al. (2014:51), a high lexical diversity value is indicative of text difficulty, since more lexical items also means more information to be processed and incorporated into the context of the text. On the other hand, low lexical density values can be a result of vocabulary repetitions and redundancies as well as low vocabulary variation, which may lead to higher text cohesion, but at the same times lowers text difficulty. Even though the GM has the highest type-token ratio – the measurement that is reportedly affected by the text length (see McNamara et al. 2014:51) – its values are the lowest in the MTLD and VOCD categories, which are based on the improved type-taken ratio measurement not affected by the text-length (McNamara et al. 2014:51). Based on these results, it can be inferred that the GM contains less information to be processed by the listener. On the other hand, having the highest values in the MLTD and VOCD categories, the FCE seems to be lexically the most diverse subtest, requiring from the listener to process much information. The C&G displays median scores in the category of lexical diversity.

The Coh-Metrix analysis also demonstrates some differences between the subtests with regard to connectives, which, according to McNamara et al. (2014), contribute to better cohesion and text ease. Even though the GM has the highest incidence of the "all connectives", the data reveal that this relatively high score can be mostly attributed to the high incidence of the simple, additive connectives (e.g., *and, also, moreover*). The FCE and the C&G, on the other hand, exhibit highest incidences of causal and logical connectives. These results may indicate that connective-wise, by resorting mostly to simple additive connectives, the GM is the closest to the authentic representation of the oral discourse (see section 3, and Buck 2001). As reported by Shohamy and Inbar (1991:23), this feature bears direct consequences for the test-takers' performance since "different types of texts located at different points on the oral/literate continuum [result] in different test scores, so that the more 'listenable' texts [are] easier." In contrast, the C&G, having the highest incidence in the category of more complex connectives (logical, adversative/contrastive and temporal connectives), resembles more the literate end of the oral/literate continuum. The high occurrence of complex connectives may also require fewer inferences from the test-takers, which may in turn reduce the overall task difficulty.

Moving to the syntactic complexity, the category of left-embeddedness in the Coh-Metrix analysis tool calculates the mean number of words preceding the main verb. Together with the mean number of modifiers per noun phrase, the high mean number implies cognitive complexity, because it requires a storage of lexical material in short term memory before it could be linked with the main verb, the main clause or the noun-head and interpreted together (McNamara et al. 2014:71). The Coh-Metrix analysis shows extremely high values in both discussed categories for the GM subtest. In fact, its left embeddedness value is more than twice higher than that of the FCE in the Q section as well as the T section, and by a half higher than that of the C&G. The values of the number of modifiers per noun phrase are more equal in the T sections, but there is again a noticeable difference in the Q section; the GM has the highest value of the three. Thus, it can be concluded that, from the perspective of syntactic complexity, the GM is by far the most challenging of the three subtests.

There are fewer differences between the three subtests with regard to the category of syntactic pattern density. According to McNamara et al. (2014:72), a high value in any of the syntactic pattern density subcategories can be indicative of a text in which the information is packed in dense syntactic structures, which consequently leads to text difficulty. Since all three subtests display very similar overall syntactic pattern density values, it can also be concluded that they are highly comparable.

The last category examines readability – the indices used are Flesch Reading Ease and the Coh-Metrix L2 Readability. Based on the Flesch Reading Ease measurement, we can observe that the Q sections are slightly more difficult to comprehend than the T sections. This is not a desirable result, since, in theory, accompanying questions should be slightly less difficult to process than the text on which the task is based. While the differences are minor in the case of the FCE and the C&G, the GM has, according to the Flesch Reading Ease measurement, significantly more demanding questions than the tapescript itself. However, examining the Coh-Metrix L2 Readability measurement, which was specially developed to predict the readability of texts for L2 students (McNamara et al. 2014:80), we can observe that the values of the three subtests are very similar, and that in all three cases the values in the Q sections are higher than those of the T sections. The latter result is also partially confirmed by the Vocaprofile analysis (Table 2) which shows that, in all three cases, Q sections feature slightly simpler vocabulary items than the T sections, so at least from the lexical perspective the Q sections are easier than the T sections.

6. Concluding Remarks and Further Research

The paper presents an analysis of syntactic and lexical complexity of three listening comprehension subtests. The study is based on Weir's socio-cognitive validation framework and it shows that, even though the test-providers claim that the tests are at the same CEFR level (B2), the analysis of some aspects of context validity demonstrates that they are not fully equivalent.

Typical of the GM are listening texts that are lexically demanding but still authentic with regard to discourse-type, i.e., they are not overly packed with information and they contain relatively many additive connectives, which is characteristic of spoken language. The GM texts have also been revealed as the most syntactically complex. The observed lexical and syntactic features may be problematic from the perspective of the number of tasks included in the exam and, consequently, the length of each task. Nevertheless, no matter what category of Vocabprofile/Coh-Metrix data is observed, the FCE rarely stands out among the three exams analysed. The sole exception is its characteristic of being informationally packed in the type-token, MTLD and VOCD categories. A similar observation can be made about the C&G – the exam is distinct from the other two only with regard to the relatively high incidence of logical, adversative/contrastive and temporal connectors, which implies that fewer inferences are required from the test-takers, and that the tapescripts are closer to literate than oral texts.

Overall, we can conclude that there is strong evidence that the GM is the most authentic of the three listening tests. The texts included in this exam are taken from live radio shows and their syntax and vocabulary are not simplified in any detectable way. The FCE and the C&G, on the other hand, appear to be scripted and place more emphasis on the variety of the texts, their informational richness and (possible) syntactic simplification

The findings also show the value of Weir's approach to validity analysis. Although the present study is limited to context validity, the results persuasively demonstrate the inadequacy of relying only on the CEFR descriptors when comparing examinations.

The discussion also indicates the need for further research: other aspects of context validity as well as other types of validity should be explored. Pertaining to the former, it would be valuable to explore speech rates since our preliminary findings indicate significant differences between the three tests: the GM has the highest average speech rate (197 words per minute) and the C&G the lowest (151 words per minute). And with regard to the latter, we believe that it would be beneficial to administer the tests to the same population, analyse the results, review CEFR alignment data, and compare the differences between B2 cut-off scores.

References:

Alderson, Charles J. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. Charles. 2006. "Analysing tests of reading and listening in relation to the common European framework of reference: The experience of the Dutch CEFR Construct Project." *Language Assessment Quarterly* 3(1):3-30.

Alderson, J. Charles, Neus Figueras, Henk Kuijper, Guenter Nold, Sauli Takala and Claire Tardieu. 2004. "The development of specifications for item development and classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening: Final report of The Dutch CEF Construct Project. Working paper" in *Lancaster University Publications & Outputs* [Online]. Available: http://eprints.lancs.ac.uk/44/1/ final_report.pdf [Accessed 2015, June 11].

American Psychological Association, American Educational Research Association and National Council on Measurement in Education. 1954. *Technical recommendations for psychological tests and diagnostic techniques*. Washington: The Association.

Bloomfield, Amber. N., Sarah C. Wayland, Allison Blodgett and Jared Linck. 2011. "Factors Related to Passage Length: Implications for Second Language Listening Comprehension" in *Expanding the Space of Cognitive Science, Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Laura Carlson, Christoph Hoelscher and Thomas. F. Shipley (Eds.). Austin: Cognitive Science Society, pp. 2317-2322.

Buck, Gary. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.

Cobb,T. n.d. "*Web Vocabprofile*, an adaptation of Heatley, Nation & Coxhead's (2002) *Range.*" [Online]. Available: http://www.lextutor.ca/vp/ [Accessed 2015, March 12].

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Heatley, A., I. S. P. Nation and A. Coxhead. 2002. "RANGE and FREQUENCY programs." [Online]. Available: http://www.victoria.ac.nz/lals/staff/paul-nation. aspx [Accessed 2015, March 12].

Ilc, Gašper and Andrej Stopar. 2015. "Validating the Slovenian national alignment to CEFR: The case of the B2 reading comprehension examination in English." *Language Testing* 32(4):443-462.

Jones, Neil. 2009. "A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard-setting" in *Linking to the CEFR levels: Research perspectives*. [Online]. Neus Figueras and José Noijons (Eds.). Arnhem: Cito, EALTA, pp. 33-43. Available: http://www.coe.int/t/dg4/linguistic/ Proceedings_CITO_EN.pdf [Accessed 2015, March 14].

Kobeleva, Polina P. 2012. "Second Language Listening and Unfamiliar Proper Names: Comprehension Barrier?" *RELC Journal* 43(1):83-98.

Matthews, Joshua and Junyu Cheng. 2015. "Recognition of high frequency words from speech as a predictor of L2 listening comprehension." *System* 52:1-13.

Messick, Samuel. 1995. "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning." *American Psychologist* 50:741-749.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

McNamara, Danielle S., Max M. Louwerse, Zhiqiang Cai and Arthur C. Graesser. 2005. *Coh-Metrix version 1.4.* [Online]. Available: http://cohmetrix.memphis.edu [Accessed April, 2015].

Rost, Michael. 1990. *Listening in Language Learning.* London: Longman.

Rost, Michael. 2006. "Areas of research that influence L2 listening instruction" in *Current Trends in the Development and Teaching of the Four Language Skills*. Esther Uso-Juan and Alicia Martinez-Flor (Eds.). New York: Mouton de Gruyter, pp. 47-74.

Shiotsu, Toshihiko and Cyril J. Weir. 2007. "The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance." *Language Testing* 24(1):99-128.

Shohamy, Elana and Ofra Inbar. 1991. "Validation of listening comprehension tests: the effect of text and question type." *Language Testing* 8(1):23-40.

Taylor, Lynda. 2004. "Issues of test comparability." *Research Notes* 15:2-5.

Weir, Cyril J. 2005. *Language testing and validation. An Evidence-Based Approach*. Houndgrave: Palgrave Macmillan.

Analysed material:

The City Guilds international examination in English
http://www.cityandguildsenglish.com/files/general.english/practice.papers/iesol/iesol.b2.communicator/iesol.b2.communicator.sample.paper.1.pdf
The First Certificate in English
UCLES. 2006. *First Certificate in English. Information for Candidates*
The General Matura in English
http://www.ric.si/mma/M141-241-1-2/2014100714441854/;
http://www.ric.si/mma/M141-241-1-4/2014100714442080/.

Notes on the authors

Gašper Ilc is an Associate Professor at the English Department, University of Ljubljana. His research interests are focussed on (formal) syntax, contrastive linguistics, and language testing. He has actively participated in international conferences on formal approaches to syntax. Since 2008, he has been President of the Testing Committee for the Matura in English (B2), which designs the examination material for the national examination in English. He has attended international conferences and workshops on language testing and assessment. His published works include papers on English syntax, contrastive linguistics, sociolinguistics, language testing, and EFL.

Andrej Stopar is an Assistant Professor at the English Department, University of Ljubljana, Slovenia. His research interests are primarily focussed on syntax, language testing, phonetics, and contrastive linguistics. The courses he has taught include English Verb, English Pronunciation, Language in Use, Translation, and Slovenian Morphology. Since 2008, Dr. Stopar has been President of the Slovenian Subject Testing Committee for the Vocational Matura in English. The group designs the tasks for the national (B1) examination in English.