VALIDITY IN HIGH- AND LOW-STAKES TESTS: A COMPARISON OF ACADEMIC VOCABULARY AND SOME LEXICAL FEATURES IN CLIL AND NON-CLIL STUDENTS' WRITTEN TEXTS

EVA OLSSON AND LISS KERSTIN SYLVÉN

[1] INTRODUCTION

In second language (L2) learning research, learners' proficiency levels and progress are often investigated. Sometimes high-stakes tests, which are part of the school curriculum, are used for this purpose, but more often tests designed for the purpose of the specific research study are utilized. How do we know that tests of the latter kind actually show what learners know and can do, when they do not have any impact on school grades? In other words, how can we be sure that our informants do as well in low-stakes tests specifically designed for research purposes as they would in high-stakes tests that result in final grades, and thus have an impact on the individual's future? The answer is, of course, that we can never know for sure. One way of finding out, though, is to compare results from high- and low-stakes tests.

In this study, we examine whether students display similar levels of performance when writing in high- and low-stakes contexts, with regard to the use of English academic vocabulary and some other linguistic features, more precisely text length, word length and variation of vocabulary. Thereby, we indirectly explore whether students have put a similar amount of effort into high- and low-stakes writing assignments. We investigate this by analyzing and comparing texts written under high- and low-stakes conditions. The purpose of the study is, firstly, to validate results obtained in the low-stakes writing assignments used in the large-scale longitudinal research project Content and Language Integration in Swedish Schools, CLISS, focusing in particular on results regarding productive academic vocabulary and the linguistic features mentioned above. Secondly, we hope that this study will shed new light on validity

in relation to writing assignments in high- and low-stakes contexts in a more general sense, for instance with regard to the role of effort and motivation.

[2] BACKGROUND

In Sweden, students attend compulsory school for nine years from the age of six or seven, and thereafter, almost all students, 98 percent, continue on to senior high school (grade levels 10-12), where a variety of educational programs are offered (the Swedish National Agency for Education, 2014). Generally, Swedish is the language of instruction in compulsory school as well as in senior high school programs, with the exception of foreign language classes. However, approximately 27 percent of all senior high schools offer educational programs where a foreign language, in most cases English, is, at least partly, the medium of instruction (Yoxsimer Paulsrud, 2014). Content and Language Integrated Learning, CLIL, is an umbrella term including a variety of teaching approaches where another language than the students' (and often also the teachers') first language (L1) is used as the medium of instruction in non-language subjects, e.g. history and biology. The main aim of CLIL is to enhance language learning, compared to what is possible in the limited time available for regular language arts classes, as subject content presented in the target language may offer authentic and linguistically challenging input and frequent opportunities to use the language (Cenoz, Genesee, & Gorter, 2014; Dalton-Puffer, 2011).

In the CLISS project, the primary aim was to study language proficiency and progress in learners' L1 Swedish and L2 English in two groups of senior high school students, aged 15/16 – 18/19. The first group consisted of students in CLIL programs, and the second group of students in parallel non-CLIL programs. Whereas the CLIL group studied several subjects through the medium of English in addition to English language arts, the non-CLIL group were instructed in Swedish and English was only studied as a separate subject. The students were followed during their three years in Swedish senior high school, and a number of tests, questionnaires, interviews and observations were administered during the time of the project (for more details on the CLISS project, see Sylvén & Ohlander, 2014). Among the tests were writing assignments in Swedish and English – in this article we will focus on the English assignments, and how the texts they generated compare with those produced in the high stakes national tests of English (NTE) with regard to academic vocabulary and certain other linguistic features.

English is considered a foreign language (FL) in Sweden, but in ongoing discussions it is debated whether in fact it should be regarded as an L2 (Hylten-

stam, 2004; Norén, 2006). In everyday Swedish society, English is encountered on a daily basis in ads, English TV- and film productions are subtitled rather than dubbed, and young people spend large amounts of their spare time involved in various activities, for instance on social media, where English is the medium of communication (Sylvén & Sundqvist, 2012; Olsson, 2012, 2016; Sundqvist & Sylvén, 2016). The large exposure to English outside school, combined with the introduction of English as a core subject in compulsory school at an early age, often already in first grade, seem to contribute to the relatively high English proficiency among Swedish students, as shown in international comparisons (European Commission/SurveyLang, 2012). It has been suggested that the use of English as the medium of instruction, as in CLIL, may not yield such overwhelmingly positive outcomes in Sweden as seen elsewhere, due to, among other things, the extramural exposure to English (Sylvén, 2013), therefore resulting in a relatively minor CLIL effect (c.f., Juan-Garau & Salazar-Noguera, 2015; Lasagabaster & Ruiz de Zarobe, 2010; Ruiz de Zarobe, 2015).

English writing proficiency is a vital skill to master for students aiming for higher education, and therefore, highly relevant to study in connection with CLIL among students enrolled in higher education preparatory programs. Of particular interest is investigating students' use and mastery of academic language (Coxhead, 2000; Hyland, 2016). As pointed out above, Swedish students' proficiency in English is generally high, but is that necessarily also true regarding academic registers? The type of language encountered outside school is more often than not of a non-academic style, more related to everyday, social language. There are some findings indicating that the effects of English encountered outside school may fade off at higher proficiency levels where academic language is needed (Olsson & Sylvén, 2015). Therefore, it is highly relevant to investigate effects of the use of English as the carrier of academic subject content. The CLISS project, of which the present study is a part, is a welcome contribution to our understanding of CLIL. Here, validation of the findings based on the writing tasks used in the CLISS project is in focus, and hopefully the study will add to our understanding of writing in high- and low-stakes contexts.

[3] LITERATURE REVIEW

In this section, research focusing on high- and low-stakes testing as well as relevant research on academic language will be accounted for.

Kane (2013) suggests that, in validation, it is the interpretations and uses of test scores, i.e. the inferences drawn from test results, which are validated. This notion is central to the present study, since we compare results from tasks

written in two different contexts. The importance of understanding and considering test-taking effort when evaluating the outcome is highlighted by Barry, Horst, Finney, Brown, and Kopp (2010). They argue that failing to take effort into account makes it difficult to know if the results actually reflect the abilities and skills of the test-taker (cf., Wise & DeMars, 2005). Test-taking effort could be expected to be higher in high-stakes assignments, which may have personal consequences for the test takers, than in low-stakes contexts, where no such consequences exist (Barry et al., 2010). Obviously, if students have put very little effort into the completion of an assignment, the actual writing proficiency will not necessarily be reflected, but rather what they can do with minimal effort. As pointed out by Eklöf (2010, p. 345), students cannot perform well without sufficient knowledge, but on the other hand, without motivation they may choose not to do their best and thus, underachieve. Hence, the validity and reliability of research results based on low-stakes assignments could be threatened, depending on what inferences are drawn from the results. Finn (2015) argues that there is growing concern that results reported without consideration of motivation will lead to biased conclusions about student knowledge.

In their overview of studies into the role of test-taker motivation and effort in test results, Wise and DeMars (2005) concluded that motivated students outperformed those who were not as motivated. In line with these results, Cole, Bergin, and Whittaker (2008) found that the perceived usefulness and importance of low-stakes tests predicted test-taking effort, and further, that the effort predicted performance. However, Barry et al. (2010) found that the cognitive demands of a test also seemed to affect the level of effort put into the assignment. While cognitive tests typically measure knowledge and skills, and items are scored as right or wrong, non-cognitive tests measure attitudes or affect, and items have no correct answer. In Barry et al. (2010) the test-takers reported lower effort for cognitively demanding tests; if the test was perceived as too difficult by the student, he or she gave up more easily.

Yet it seems possible for test administrators to strengthen student motivation even in low-stakes contexts. Wise and DeMars (2005, p. 15) argue that “[l]ow-stakes assessment testing can yield scores that are valid, as long as we accept the responsibility for effectively managing student motivation”. Findings by Attali (2016) indicate that test takers in low-stakes settings may perform at the same level as in high-stakes settings with relatively little effort. Hence, it seems that the relationship between high- and low-stakes tests is not dichotomous; rather, it could be described as a multifaceted continuum where a large number of factors influence the outcome in high- as well as low-stakes contexts.

When writing proficiency under high- and low-stakes conditions is compared, as in the present study, the design of the writing assignments is undoubtedly of great importance to the outcome, as two completely different tasks would most likely result in very different texts, regardless of whether they were written in high- or low-stakes contexts. In a large-scale study of writing task variants by Bridgeman, Trapani and Bivens-Tatum (2011), all participants wrote two different variants. The task variants were related and could, for instance, introduce similar content but request the students to discuss or explain various aspects of the content. The results revealed that students' texts based on such task variants tended to be rated at similar levels. In other words, the outcome of the tasks was comparable with regard to rater's judgement of the quality of the texts.

The standard setting writing tasks included in the NTE in Sweden are designed in a similar manner, where students at senior high school are asked to argue, discuss, explore or explain certain topics. Each year, the tasks are similar in design, and as tasks are piloted and benchmarks provided, the scores for rating tend to be stable over time (for examples of tasks, see <http://nafs.gu.se>; see also Erickson, 2010). In the present study, students' texts based on a writing assignment included in the NTE are compared to the same students' texts produced in connection with assignments administered within the CLISS project (see section on methods and material).

The topics and suggested text types of all writing assignments included in this study invited the students to use academic language. Writing academic texts may be challenging for students, not least for L2 writers, as language use in academic contexts may differ substantially from language use in other contexts, such as speaking in everyday situations or when writing narratives (Cummins, 1979, 2008; cf., Schleppegrell, 2004). In research on register variation, that is, language choices made in different contexts, corpus-based methods are often used, as typical linguistic features of certain registers may be identified in corpora covering material from different contexts. Using such methods, Biber (2009) found, for example, a greater diversity of vocabulary in written university registers than in spoken registers, and further, that a larger number of nouns and nominalizations were used in writing.

As words are used to build language, there is an obvious connection between vocabulary knowledge and writing proficiency. Laufer and Nation (1995) found that the vocabulary size of L2 writers predicted how successful their written production was. Research findings show that scores for lexical measures, for instance vocabulary size and range, tend to correlate with holistic scores of writing quality (e.g., Crossley, Salsbury, & McNamara, 2012; Grant & Ginther,

2000; Hinkel, 2011). It has been shown that longer texts are often judged as better than shorter texts (Grant & Ginther, 2000). Further, scores for vocabulary variation and average word length in texts also tend to correlate with holistic assessment scores (Crossley & McNamara, 2012). In the present study, comparisons of text length, average word length and variation of vocabulary are made between texts written under high- and low-stakes conditions.

However, the main focus of the text analysis conducted here is students' use of academic vocabulary, and more precisely, their use of general academic vocabulary. In contrast to domain-specific vocabulary, consisting of content-specific words used in certain contexts, for instance in medicine or archaeology, general academic vocabulary appears across different disciplines but not as frequently in non-academic contexts (Baumann & Graves, 2010). Various academic word lists have been compiled from text corpora, such as the Academic Word List (AWL; Coxhead, 2000) and the more recent Academic Vocabulary List (AVL), compiled by Gardner and Davies (2014). Different methods were used in the compilation of these lists, which, naturally, have implications for the vocabulary items included in the respective list. The AVL, which is more extensive and has a larger coverage than the AWL, was used in this study. (For a comparison of analyses of CLISS students' texts based on the AWL and the AVL, see Olsson, 2015.)

The AVL is based on lemmas, that is, individual words plus inflections, and contains 3000 lemmas. The AVL was compiled from the academic section of the Corpus of Contemporary American English, the COCA (Davies, 2012), which includes more than 120 million words out of the total of 425 million words in the COCA. The academic section comprises texts which were published in the USA, covering nine disciplines.

To appear in the AVL, a word needs to appear at least 50 percent more frequently in the academic corpus than in the non-academic part of the COCA and has to occur in at least seven out of the nine disciplines, that is, across domains (for details, see Gardner & Davies, 2014). The AVL covers 13.8 percent of the words in the academic section of the COCA, on which it was built, and 13.7 percent of the academic section of the British National Corpus (BNC, Nation, 2004). In this study, the vocabulary used by the students is compared to the AVL to investigate the extent to which students use general academic vocabulary.

[4] METHOD AND MATERIALS

The main purpose of this study, as mentioned above, is to validate results obtained in the low-stakes English writing assignments used in the CLISS project.

In the large scale CLISS project, CLIL and non-CLIL students' ($N = 245$) L2 English writing proficiency and progress were investigated. To that end, writing assignments were administered to the students four times during the three-year period of the project. These assignments were written under low-stakes conditions; they were not an obvious part of regular school work or even marked by the teachers. Nevertheless, all CLISS assignments were designed to mirror the type of writing assignments included in the NTE, which are high-stakes tests for the students, as test results are likely to impact their final grades (see The Swedish National Agency for Education: www.skolverket.se).

A crucial issue when examining whether the assignments used in the CLISS project have provided valid data is to consider the purpose of collecting the texts, in other words, the objective of the study. The CLISS assignments were designed for the purpose of eliciting texts, where students' progress in academic language could be investigated, with particular attention paid to academic vocabulary. The topics of the assignments related to the natural and the social sciences, school subjects studied by all of the participants, and the suggested text types were typical of school-related writing at senior high school, as students were asked to write exploratory and argumentative texts. The following assignments were given:

- (i) For or against nuclear power (argumentative text)
- (ii) Matters of gender and equality (expository text)
- (iii) Ways to political and social change – violence or non-violence (argumentative text)
- (iv) Biodiversity for a sustainable society (expository text)

Written instructions accompanied each assignment, including one or two pages of factual texts, diagrams or pictures for inspiration. The texts were written at school, using computers. The time limit for the two first assignments was 120 minutes, but as all students finished well within 90 minutes, the two last assignments were reduced to 90 minutes. The first task was given shortly after the students started in senior high school in grade level 10; thus, results from that task are to be considered as baseline data. The second and third CLISS assignments were given in grade level 11 and the last assignment in the final year, grade level 12.

The high-stakes assignment analysed in the present study is a writing assignment given as a part of the NTE in grade level 11, and after the third CLISS

assignment. The writing assignments included in the NTE are reused a number of times and as the instructions for the particular assignment used in this study constitute classified information at this point in time, they cannot be shared. However, as already mentioned, in the NTE at senior high school level, students are asked, for instance, to take a stand on a controversial subject or to explore a topic often related to life in today’s society (for examples of tasks, see <http://nafs.gu.se>). Normally, students are advised to write a text of between 300 and 600 words, and they are generally allowed 100 minutes for the task.

The CLISS assignments were designed to mirror those given in the NTE with regard to topics and suggested text types. However, no indication of text length was given in the CLISS assignments, only a time limit. Further, the instructions for the CLISS assignments were somewhat more extensive compared to those for the NTE assignment. As the content covered by the CLISS assignments clearly related to the natural or the social sciences, some graphs and numbers were given to enable all students to write a text even if their detailed knowledge of, for instance, nuclear power was poor. In a similar vein, the NTE assignments also provide background information, but often in the form of a short introductory text followed by supporting questions or mind maps for inspiration.

In the present study, texts written by 66 students (33 CLIL and 33 non-CLIL) at one of the schools taking part in the CLISS project were used since access was granted to the NTE texts at this school. As some students changed schools or were absent on certain occasions, the number of students who completed each assignment varied over time. In all, 221 texts were included in the study. In Table 1, the number of texts per assignment and group (CLIL/non-CLIL) is shown in chronological order.

Assignment	CLIL N	Non-CLIL N	Total N
CLISS 1	24	19	43
CLISS 2	21	18	39
CLISS 3	24	25	49
National Test	27	28	55
CLISS 4	23	12	35
Total	119	102	221

TABLE 1. The number of texts by CLIL and non-CLIL students per assignment.

The main issue investigated here is whether the CLISS assignments seem to have elicited academic vocabulary use to a similar extent as the NTE even

though they were written in low-stakes contexts. Some other text features – text length, average word length and variation of vocabulary – are also compared, as these features may also indicate whether students' efforts seem to have been at a similar level when writing the low-stakes assignments as when writing the high-stakes national test. Since it was beyond the scope of the CLISS project to distribute surveys of motivation in direct connection to the writing assignments, there are no such data available. Instead, we have analysed the same type of linguistic features in texts written in both the high- and low-stakes contexts to establish if they result in similar outcomes with regard to the features in focus.

[4.1] *Method of text analysis*

The vocabulary used by the students was compared to the Academic Vocabulary List (Gardner & Davies, 2014), and the proportion of vocabulary covered by the list was noted for each of the texts. This analysis was based on tokens, that is, a word was counted each time it occurred. For the analysis of academic vocabulary, an interface available at <http://www.wordandphrase.info/academic/> was used.

Further, the length of each text, as the number of running words, was noted, as was the average word length (in characters) of each text. The variation of vocabulary in the texts was analysed using a type/token ratio, where the number of different words in a text is divided by the total number of words. As shorter texts tend to result in higher ratios than longer ones, a standardised type/token ratio was used, where the ratio is calculated every 100th word and then the average ratio for the text. By using the standardised type/token ratio, differences in text length are controlled for. For these analyses, WordSmith Tools 5.0 (<http://www.lexically.net/wordsmith>) was used.

Statistical analyses were conducted comparing scores for the CLISS assignments to scores for the NTE. In addition, comparisons between the two groups of students involved, CLIL and non-CLIL, were made. For statistical analyses, independent samples t-tests as well as paired samples t-tests were run, using SPSS, version 24.

[5] RESULTS

In Table 2, the scores for academic vocabulary, calculated as the percentage of vocabulary covered by the Academic Vocabulary List (AVL; Gardner & Davies, 2014), are shown for the low-stakes assignments, the four CLISS assignments, and for the high-stakes assignment, the NTE. The results in Table 2 are given in

chronological order. AVL scores are shown for CLIL and non-CLIL groups, as are the results of statistical comparisons between the two groups.

Assignment	CLIL			Non-CLIL			t	p
	N	AVL% Mean	SD	N	AVL% Mean	SD		
CLISS 1	24	6.6	3.1	19	5.7	2.9	0.86	.393
CLISS 2	21	9.7	4.2	18	7.7	3.4	1.58	.124
CLISS 3	24	8.5	3.5	25	6.0	2.7	2.80	.007*
National Test	27	8.7	3.9	28	7.3	3.1	1.52	.133
CLISS 4	23	12.9	3.9	12	10.6	3.8	1.70	.099

TABLE 2. The proportion of academic vocabulary (AVL%) in texts by CLIL and non-CLIL students.

Table 2 shows that both CLIL and non-CLIL students increased their use of academic vocabulary throughout the three years of senior high school (grade levels 10-12); the first CLISS assignment included a smaller proportion of academic vocabulary than the last. There was a dip in the third assignment, but otherwise, the use of academic vocabulary progressed. As indicated in Table 2, an independent samples t-test revealed that there was no statistically significant difference in the use of academic vocabulary between CLIL and non-CLIL students, except in assignment 3, where CLIL students used a significantly larger proportion of academic vocabulary. (For comparisons between CLIL and non-CLIL students' use of academic vocabulary in the CLISS assignments based on all 245 students participating in the project, see Olsson, 2015; 2016; Olsson & Sylvén, 2015).

The mean scores shown in Table 2 indicate that the low-stakes CLISS assignments seem to have elicited academic vocabulary to the same extent as the high-stakes NTE, regardless of whether the students belonged to the CLIL or non-CLIL group. On average, the students used a larger proportion of academic vocabulary in the NTE than in CLISS assignments 1 and 3, both written at an earlier point in time, but a smaller proportion than in CLISS assignment 2, which was also written at an earlier point in time. As students are likely to increase their academic vocabulary throughout their education, the higher scores exhibited in CLISS assignment 4 were to be expected. Other factors, such as the topic of the assignment or the requested text type may also have had an influence on the use of academic vocabulary, which may explain the discrepancy between CLISS assignments 2 and 3 (cf., e.g. Beers & Nagy, 2009).

However, to further compare the outcome of the low-stakes CLISS assignments to that of the high-stakes NTE, paired analyses were made. In other words, a student's score for AVL in the NTE was compared to his or her scores in each of the CLISS assignments. The results show the degree of correlation between the rank order of students, based on their scores for AVL in the NTE and each of the CLISS assignments. If, for instance, the student with the highest score in the NTE also had the highest score in the CLISS assignment, the scores correlate. Further, the paired t-test reveals if there are statistically significant differences in the comparison of means (AVL%) between the NTE and each of the CLISS assignments. In the paired analysis, no division was made into CLIL and non-CLIL groups as the purpose here was to compare production in high- and low-stakes assignments.

The paired samples t-tests show that there are statistically significant correlations between the proportion of AVL vocabulary in the NTE and each of the four CLISS assignments (CLISS 1: $r = .67$ $p < .001$; CLISS 2: $r = .41$ $p = .010$; CLISS 3: $r = .63$ $p < .001$; CLISS 4: $r = .70$ $p < .001$). This means that the rank order of students, based on their scores for AVL, correlates in a statistically significant way between the NTE and each of the CLISS assignments. In other words, students who scored high (or low) in the NTE tended to do so in the CLISS assignments as well.

Turning to the comparison of mean scores for AVL in the NTE and the CLISS assignments, the paired t-tests revealed that there was no statistically significant difference between the NTE and CLISS 2 or CLISS 3. The students used a significantly larger proportion of academic vocabulary in CLISS 1 and CLISS 4 than in the NTE (CLISS 1: $t = 3.93$ $p < .001$; CLISS 4: $t = 7.18$ $p < .001$). This means that the NTE did not elicit academic vocabulary to a greater extent than any of the CLISS assignments. On the contrary, two of the CLISS assignments, the first and the last, elicited a significantly larger proportion of academic vocabulary. Thus, the students used academic vocabulary at least to the same extent in the low-stakes CLISS assignments as in the high-stakes NTE, which strengthens the validity of the CLISS assignments for the purpose of measuring productive academic vocabulary.

In addition to the analysis of academic vocabulary, there are, of course, many other text features that may be of great relevance to study when comparing the outcome of high- and low-stakes assignments. Here, we have analysed some other linguistic features that often correlate with a holistic judgment of text quality. In Table 3, the average scores for text length (in number of running words), for word length (in number of characters) and for variation of vocabulary (measured in a standardized type/token ratio) are shown for the NTE

and for the four CLISS assignments.

	Text length Mean	SD	Word length Mean	SD	Type/token Mean	SD
CLISS 1	506.4	202.5	4.3	.2	67.1	2.7
CLISS 2	413.0	144.4	4.3	.3	66.3	4.7
CLISS 3	463.7	201.2	4.4	.3	67.7	4.4
National Test	476.6	100.8	4.5	.3	69.5	3.8
CLISS 4	474.5	145.8	4.5	.3	67.8	3.5

TABLE 3. Mean text length, word length and variation of vocabulary in CLISS assignments and the National Test.

Paired samples t-tests show that there was a statistically significant correlation of text length between the NTE and CLISS 2 ($r = 0.40$ $p = .014$) and NTE and CLISS 3 ($r = .34$ $p = .016$). The results reveal that students who wrote long (or short) texts in the NTE tended also to do so in CLISS 2 and 3; the rank order of the students based on the length of their texts correlated in a statistically significant way. However, there was no statistically significant correlation in text length between the NTE and CLISS 1 or 4.

Turning to the comparison of mean scores for text length in the NTE and the CLISS assignments, the paired t-tests revealed that there were no statistically significant differences between the NTE and any of the CLISS assignments in this respect. Thus, with regard to text length, the high- and low-stakes assignments seem to have generated similar outcomes.

As regards word length, paired samples t-tests showed that there was quite a strong correlation between the NTE and all the CLISS assignments (CLISS 1: $r = .74$ $p < .001$; CLISS 2: $r = .60$ $p < .001$; CLISS 3: $r = .61$ $p < .001$; CLISS 4: $r = .74$ $p < .001$). In other words, the students who tended to use long (or short) words in the NTE, tended also to do so in the CLISS assignments. Thus, the rank order of students, based on the average word length in their texts, correlates in a statistically significant way between the NTE and each of the other assignments.

The paired analysis also revealed that there were statistically significant differences in average word length between the NTE and three of the CLISS assignments (CLISS 1: $t = 6.09$ $p < .001$; CLISS 2: $t = 4.85$ $p < .001$; CLISS 3: $t = 3.47$ $p = .001$). There was no statistically significant difference in word length between the NTE and CLISS 4. These results show that the students tended to use longer

words on average in the NTE than in the CLISS assignments written before the NTE, but that the last CLISS assignment included words of a similar average length as the NTE did.

Turning to the variation of vocabulary, measured in a standardised type/token ratio, the paired samples t-tests revealed that there are statistically significant correlations between the NTE and the four CLISS assignments (CLISS 1: $r = .38$ $p = .016$; CLISS 2: $r = .52$ $p = .001$; CLISS 3: $r = .41$ $p = .003$; CLISS 4: $r = .37$ $p = .033$). Thus, the rank order of students, based on the variation of vocabulary in their texts, correlates in a statistically significant way between the NTE and each of the CLISS assignments.

The paired t-tests also indicated statistically significant differences in the variation of vocabulary between the NTE and the CLISS assignments (CLISS 1: $t = 4.54$ $p < .001$; CLISS 2: $t = 4.41$ $p < .001$; CLISS 3: $t = 2.68$ $p = .010$; CLISS 4: $t = 3.92$ $p < .001$). Clearly, the students tended to use more varied vocabulary in the NTE than in the low-stakes CLISS assignments. However, as already mentioned, the students used academic vocabulary to a greater extent in CLISS 4 than in the NTE.

[6] DISCUSSION

This study set out to validate results obtained in low-stakes assignments by comparing them to those obtained in comparable high-stakes tests. The findings show that students seem to have used academic vocabulary to the same extent when writing the low-stakes CLISS assignments as when writing the high-stakes NTE. The topics of the CLISS tasks were clearly related to the curricula of the natural or the social sciences, more so than the NTE, which may explain why they contained a rather high proportion of academic vocabulary in spite of the low-stakes conditions. Even though the two types of assignment were slightly different as regards topic, the CLISS assignments closely mirrored the NTE, so this should not have had a major effect on comparability. If domain-specific vocabulary had been the objective of the analysis, the validity of the comparison between the CLISS assignments and the NTE could have been threatened, but as it was, general academic vocabulary was in focus. The results revealed that general academic vocabulary was used at least to the same extent in the CLISS assignments as in the NTE – in the last of the CLISS tasks the proportion of academic vocabulary is actually higher than in the NTE. Hence, we conclude that the validity of the CLISS assignments, for the purpose of eliciting text where progress in academic vocabulary can be studied, is strengthened by the comparison with the NTE.

The strengthened validity, however, does not automatically lead to the con-

clusion that the students were as motivated to perform well when writing the CLISS assignments as when writing the NTE. Some significant differences were found where the NTE generated higher scores for features such as variation of vocabulary, which could indicate that student effort was greater. It is very likely that a high-stakes exam is considered much more important than assignments written for research, as shown, for instance, by Wise and DeMars (2005). On the other hand, too much pressure may lead to lower achievement (Embse & Hasson, 2012; Putwain, 2008). CLIL students, in particular, might have aimed at performing well, as they had chosen a program where English was used as a language of instruction. Findings from a study by Sylvén and Thompson (2015), comparing levels of motivation among the students involved in the CLISS project, revealed that the CLIL students were more highly motivated to learn and use English than non-CLIL students. This might have made the CLIL students more motivated to do their best when writing the assignments (c.f., Thompson & Sylvén, 2015; Henry, Davydenko & Dörnyei, 2015).

As pointed out in the introduction of this article, the validity of writing assignments and the conditions under which they were written will strongly influence the validity of results. In a longitudinal study where different tasks and topics are used, it may be difficult to separate time- and task-induced variability (Ortega & Iberri-Shea, 2015). Each additional school year could be expected to result in enhanced writing proficiency, provided that writing is practised, and hence, scores could be expected to increase over the three years of the study. However, in the text analyses in the present study, not only the chronological order of the assignments seemed to influence scores; for instance, the first CLISS assignment produced longer texts on average than the other assignments. The third CLISS assignment generated a more limited use of academic vocabulary than the preceding assignment. Speculatively, the topic, *Ways to political and social change – violence or non-violence*, was too difficult for 17-year-olds, especially as some of the individuals mentioned in the background material, for instance Che Guevara, may not have been known to them. Even if using the same or very similar tasks in all assignments could have diminished the risk of topic-induced variability, this could have been extremely demotivating for students and thus, the validity of results would have been more severely threatened. Further, a practice effect, that is change in performance due to repetition, may also occur when the same task is used several times (Barkaoui, 2014; cf., Gustafsson, 2010).

The focus of this study has been comparisons of assignments written in high- and low-stakes contexts; let us now briefly comment on the comparison between CLIL and non-CLIL students, as these two groups provided the texts

that were analysed. The results show that CLIL students in the present study did not use academic vocabulary to any significantly greater extent than non-CLIL students, either in three of the four CLISS assignments or in the NTE. When texts written by students from all of the three schools involved in the CLISS project were included in the analysis, it was found that CLIL students did not progress more than non-CLIL students in their use of academic vocabulary over the three years, when initial differences were controlled for (Olsson 2015, Olsson 2016). These results were to some extent unexpected since CLIL students more often encounter and use English in lessons and literature where academic language would be expected. Further research into different ways of supporting the development of academic language in CLIL is therefore called for. Also, further analyses of the qualitative aspects of academic language in the texts used in the CLISS project are required, as only the occurrence of academic vocabulary and certain other text features, for instance text length, were studied here.

It is not possible to draw any general conclusions about the validity of results based on low-stakes assignments or the comparability of results from high- and low-stakes assignments based on the present limited study. However, the results of this study may illustrate that a comparison of certain textual features will strengthen the validity of results based on low-stakes assignments with regard to those particular features, in line with Kane (2013). In the present study, academic vocabulary was analysed in the texts, and the results revealed that the high- and low-stakes assignments used here seemed comparable in this respect, but not with regard to, for instance, variation of vocabulary. The results of the present study confirm the position taken by others, e.g. Barry et al. (2010), namely that when using low-stakes assignments in research, careful consideration must be given as to whether the results actually reflect the abilities and skills of the test-taker, or if other factors are at work. We argue that the methods applied in this study constitute one way of doing this.

REFERENCES

- Attali, Yigal. 2016. Effort in Low-Stakes Assessments: What Does It Take to Perform as Well as in a High-Stakes Setting? *Educational and Psychological Measurement*, 76(6), 1045-1058.
- Barkaoui, Khaled. 2014. Quantitative approaches for analyzing longitudinal data in second language research. *Annual Review of Applied Linguistics* (34), 65-101.
- Barry, Carol L., S. Jeanne Horst, Sara J. Finney, Allison R. Brown & Jason P. Kopp. 2010. Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, 10(4), 342-363.
- Baumann, James F. & Michael F. Graves. 2010. What is academic vocabulary?. *Journal of Adolescent & Adult Literacy*, 54(1), 4-12. doi:10.1598/jaal.54.1.1
- Beers, Scott F. & William E. Nagy. 2009. Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre?. *Reading and Writing*, 22(2), 185-200.
- Biber, Douglas. 2009. Are there linguistic consequences of literacy? Comparing the potentials of language use in speech and writing. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy*. New York: Cambridge University Press.
- Bridgeman, Brent, Catherine Trapani & Jennifer Bivens-Tatum. 2011. Comparability of essay question variants. *Assessing Writing*, 16(4), 237-255.
- Cenoz, Jasone, Fred Genesee & Durk Gorter. 2014. Critical analysis of CLIL: Taking stock and looking forward. *Applied Linguistics*, 35(3), 243-262. doi:doi:10.1093/applin/amt011
- Cole, James S., David A. Bergin & Tiffany A. Whittaker. 2008. Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, 33(4), 609-624.
- Coxhead, Averil. 2000. A new academic word list. *TESOL*, 34(2), 213-238.
- Crossley, Scott A. & Danielle S. McNamara. 2012. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 75(2), 115-135.

- Crossley, Scott A., Tom Salsbury & Danielle S. McNamara. 2012. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243-263.
- Cummins, Jim. 1979. Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, 19, 198-203.
- Cummins, Jim. 2008. BICS and CALP: Empirical and theoretical status of the distinction. In B. Street & N. H. Hornberger (Eds.), *Encyclopedia of language and education*. Vol. 2, *Literacy* (2nd ed.), 71-83. New York, NY: Springer.
- Dalton-Puffer, Christiane. 2011. Content-and-Language Integrated Learning: From Practice to Principles? *Annual Review of Applied Linguistics* (31), 182-204.
- Davies, Mark. 2012. *Corpus of contemporary American English (1990-2012)* (Publication no. <http://corpus.byu.edu/coca/>).
- Eklöf, Hanna. 2010. Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, 17(4), 345-356.
- Embse, Nathaniel V. D. & Robert Hasson. 2012. Test anxiety and high-stakes test performance between school settings: Implications for educators. *Preventing School Failure: Alternative Education for Children and Youth*, 56(3), 180-187.
- Erickson, Gudrun. 2010. Good practice in language testing and assessment: A matter of responsibility and respect. In Kao, T. and Lin Y. (Eds.), *A new look at teaching and testing: English as subject and vehicle*, 237-258. Taiwan: Bookman Books.
- European Commission/SurveyLang. 2012. *First European Survey on Language Competences: Final Report*. Retrieved from <http://ec.europa.eu/languages/eslc/index.html>
- Finn, Bridgid. 2015. Measuring Motivation in Low-Stakes Assessments. *ETS Research Report Series* (2), 1-17.
- Gardner, Dee & Mark Davies. 2014. A New Academic Vocabulary List. *Applied Linguistics*, 35(3), 305-327. doi:10.1093/applin/amt015
- Grant, Leslie & April Ginther. 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123-145.

- Gustafsson, Jan-Eric. 2010. Longitudinal designs. In B. P. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research*. 77-101. New York: Routledge.
- Henry, Alastair, Sofia Davydenko & Zoltan Dörnyei. 2015. The anatomy of Directed Motivational Currents: Exploring intense and enduring periods of L2 motivation. *The Modern Language Journal*, 99(2), 329-345.
- Hinkel, Eli. 2011. What research on second language tells us and what it doesn't. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Vol. 2. New York: Routledge.
- Hyland, Ken. 2016. General and specific EAP. In K. Hyland & P. Shaw (Eds.), *Routledge Handbook of EAP*. 17-29. London: Routledge.
- Hyltenstam, Kenneth. 2004. Engelskan, skolans språkundervisning och svensk språkpolitik. In B. Lindgren & O. Josephson (Eds.), *Engelskan i Sverige*. 36-107. Stockholm: Svenska Språknämnden.
- Juan-Garau, Maria & Joana Salazar-Noguera. 2015. Learning English and learning through English: insights from secondary education. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual educational environments*, 105-122. Heidelberg: Springer.
- Kane, Michael T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Lasagabaster, David & Yolanda Ruiz de Zarobe. (Eds.). 2010. *CLIL in Spain: Implementation, Results and Teacher Training*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Laufer, Batia & Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307-322.
- Nation, Paul. 2004. A study of the most frequent word families in the British National Corpus. In P. Boogards & B. Laufer (Eds.), *Vocabulary in a second language*, 3-13. Amsterdam: John Benjamins
- Norén, Kerstin. 2006. Engelska är bättre, ibland. *Språkvård*, 3, 35-39.
- Olsson, Eva. 2012. "Everything I read on the Internet is in English". *On the impact of extramural English on Swedish 16-year-old pupils' writing proficiency*. Licentiate thesis. Gothenburg: University of Gothenburg.

- Olsson, Eva. 2015. Progress in English academic vocabulary use in writing among CLIL and non-CLIL students in Sweden. *Moderna Språk* (2), 51-74.
- Olsson, Eva. 2016. *On the impact of extramural English and CLIL on productive vocabulary*. PhD thesis. Gothenburg: University of Gothenburg Retrieved from <https://gupea.ub.gu.se/handle/2077/41359>
- Olsson, Eva & Liss Kerstin Sylvé. 2015. Extramural English and academic vocabulary. A longitudinal study of CLIL and non-CLIL students in Sweden. *Apples - Journal of Applied Language Studies*, 9(2), 77-103.
- Ortega, Lourdes & Gina Iberri-Shea. 2015. Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics* (25), 26-45.
- Putwain, David William. 2008. Test anxiety and GCSE performance: The effect of gender and socio-economic background. *Educational Psychology in Practice*, 24(4), 319-334.
- Ruiz de Zarobe, Yolanda. 2015. The effects of implementing CLIL in education. In M. Juan-Garau & J. Salazar-Noguera (Eds.), *Content-based language learning in multilingual education environments*, 51-68. New York: Springer International Publishing.
- Schleppegrell, Mary. 2004. *The language of schooling; a functional linguistics perspective*. London, UK: Lawrence Erlbaum Associates.
- Sundqvist, Pia & Liss Kerstin Sylvé. 2016. *Extramural English in Teaching and Learning: From Theory and Research to Practice*. London: Palgrave Macmillan.
- Sylvé, Liss Kerstin. 2013. CLIL in Sweden – Why does it not work? A metaperspective on CLIL across contexts in Europe. *International Journal of Bilingual Education and Bilingualism*, 16(3), 301-320.
- Sylvé, Liss Kerstin & Sölve Ohlander. 2014. The CLISS project: Receptive vocabulary in CLIL versus non-CLIL groups. *Moderna Språk*, 108(2), 80-114. doi:<http://ojs.ub.gu.se/ojs/index.php/modernasprak/article/view/2789/2614>
- Sylvé, Liss Kerstin & Pia Sundqvist. 2012. Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, 24(3), 302-321. doi:10.1017/S095834401200016X
- Sylvé, Liss Kerstin & Amy S. Thompson. 2015. Language learning motivation

and CLIL: Is there a connection? *Journal of Immersion and Content-Based Language Education*, 3(1), 28-50.

The Swedish National Agency for Education 2014. Statistik och utvärdering [Statistics and evaluation]. Accessed 19 August 2017 at <https://www.skolverket.se/statistik-och-utvardering/nyhetsarkiv/nyheter-2014/nastan-alla-grundskoleelever-fortsatter-till-gymnasieskolan-1.223182>

Thompson, Amy S. & Liss Kerstin Sylvén. 2015. Does English make you nervous? *Apples - Journal of Applied Language Studies*, 9(2), 1-23.

Wise, Steven L. & Christine E. DeMars. 2005. Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, 10(1), 1-17.

Yoxsimer Paulsrud, BethAnne. 2014. *English-medium instruction in Sweden. Perspectives and practices in two upper secondary schools*. PhD thesis. Stockholm: Stockholm University.

CONTACTS

Eva Olsson
University of Gothenburg
eva.olsson@ped.gu.se

Liss Kerstin Sylvén
University of Gothenburg
lisskerstin.sylvén@ped.gu.se