Using DIALANG to track English language learners' progress over time

Nantia Kektsidou
Department of English Language Studies, Faculty of Humanities,
University of Cyprus, Cyprus

Dina Tsagari
Department of Primary and Secondary Teacher Education,
Faculty of Education and International Studies,
OsloMet - Oslo Metropolitan University, Oslo, Norway

The current paper presents a case study where an attempt was made to re-purpose an existing language testing system. More specifically, the paper investigated the suitability of an online diagnostic assessment system (DIALANG) for tracking the English language proficiency of three groups of university students in four different skills/aspects over a period of three years. The study presents the results, issues and challenges that arise when an assessment system is used for a different purpose than it was originally designed for and discusses the lessons learnt.

Key words: second language testing, DIALANG, diagnosis, validity, test retrofitting

Introduction

Language tests are traditionally categorized into achievement, progress, placement, proficiency and diagnostic. Scores generated from language tests are used in decision making processes, which in turn define the purpose of a test (Fulcher & Davidson, 2009). Commonly, language tests are used for the purpose they were originally designed for but there are instances where a re-focus of the test is needed. An example of this is test purpose alterations referred to as 'a change retrofit', where "the test is altered to meet a completely new purpose for which it was not originally intended, or to be used with users who were not envisaged in the original statement of test purpose" (Fulcher & Davidson, 2009, p. 124). In this paper, we focus on the evaluation of the change retrofit (not alterations in test design) of an online test (DIALANG)

Email address for correspondence: nkektsidou@gmail.com

designed to diagnose language skills and meant for individual self-assessment, for the purpose of tracking the development of the English language skills of university students over time. The results and lessons learnt from such an endeavour will be of interest to the international readership working on language assessment and can shed light on issues that arise when using an assessment system developed for a different purpose than that originally intended.

Literature review

English for academic purposes (EAP) entails training students in higher education institutions to use language appropriately for academic study. EAP is one of the most common forms of English for specific purposes (ESP). An EAP program bases instruction on skills required to perform in an English-speaking academic context and on various core subject areas generally encountered in a university setting. Programmes usually include a narrow focus on the more specific linguistic demands of a particular area of study, for example business subjects. Such programs may be divided into 1) pre-sessional courses (quite common in UK universities) and 2) courses taken alongside students' other subjects. These in-sessional courses are designed to help students develop their language skills and academic practices. In case of a language focus, EAP instruction often teaches grammar, vocabulary and the four skills (reading, writing, listening and speaking – pronunciation included), but usually tries to tie these to the specific study needs of the students (Douglas, 2000).

There is some debate amongst EAP teachers as to the best way to assess students' academic English levels. This is mainly due to lack of assessment tools that can help teachers and institutions gauge students' learning that takes place as part of EAP programmes. As a result, teachers would resort to available sources such as diagnostic tests that are freely accessible. Diagnostic tests are usually realized within the domain of achievement or proficiency tests (Bachman, 1990; Mousavi, 2002; Shohamy, 1992) and quite frequently practitioners in the field have utilized the terms 'diagnostic' and 'placement' tests interchangeably (Brown, 2004; Davies, 1999). Nevertheless, recent research conducted outside the field of second/foreign language testing established that a common definition of diagnosis involves 'a formal activity of evaluation and judgement, which focuses on problem identification, and sometimes problem-solution or management, and which tends to focus more on weaknesses than strengths' (Alderson et al., 2015a, p. 254). Therefore, diagnostic tests are designed with the aim to provide users with diagnostic information about their abilities in a foreign language. The following section will present the online diagnostic instrument that was used in the current study and will review research that has so far been conducted, followed by a description of the local educational system where the study was carried out.

The DIALANG testing system

DIALANG (https://dialangweb.lancaster.ac.uk/) was designed to help learners self-assess their abilities and obtain diagnostic information about their language proficiency. It is an open-access platform, available in fourteen European languages and 18 instructional and feedback languages (see Appendix A for examples). With 250,000 tests taken in 2017, DIALANG continues to be a popular means of language assessment (Harding et al, 2018). The test was developed by an expert team at the University of Lancaster and was carried out with the financial support of the European Commission, Directorate-General for Education and Culture (SOCRATES Programme, LINGUA Action D; Klimova & Hubackova, 2013). According to Alderson and Huhta (2005), DIALANG is the first large-scale language assessment system that aims at diagnosing rather than certifying language proficiency. It does not provide a 'clinical diagnosis' of learning problems, neither is it related to specific language courses or curricula (Huhta & Figueras, 2004). Rather, its main purpose is to inform learners about their strengths and weaknesses in the following five language skills: reading, listening, writing, structures and vocabulary. Writing is tested indirectly and the tasks are similar to those in the vocabulary, reading and structures subcomponents where the requirement is to produce only a few words instead of an extensive piece of writing. The skill of speaking is not included in the web-based system of DIALANG since, according to the designers of the test, "automatic scoring of speaking and writing is very difficult and expensive to develop" (https://www.lancaster.ac.uk/researchenterprise/dialang/about).

The system also includes an optional vocabulary size placement test (VSPT) and a self-assessment test, both of which can be taken prior to the administration of the language test. The VSPT measures the users' vocabulary knowledge in order to place them at an appropriate level and then provides them with the most suitable versions of the test based on their abilities (see Appendix C). The same applies to the self-assessment test which is available only for the reading, writing and listening skills. According to Alderson and Huhta (2005), 'the VSPT was introduced to the system because there was uncertainty about the adequacy of self-assessment alone to serve as a pre-estimator of proficiency that would be used to take the user to the most suitable level of the test' (p. 303-304). Therefore, both the VSPT and the self-assessment test determine the level of the language test in which users will be placed.

Three different levels of tests for each skill are available (easy, intermediate, and difficult). These are administered automatically by the system since DIALANG is a computer-adaptive test and learners cannot self-select the level. Moreover, the VSPT and self-assessment are optional, thus, learners can omit them if so they wish. However, if they skip both tests, the system automatically administers the intermediate level of the skill they select to be tested on.

The specifications of the tests are based on the Common European Framework of Reference (Council of Europe, 2001) and the results are reported back to the learner on the six-level scale of the CEFR, from A1 to C2 (A1 being the lowest and C2 being the highest; see Appendix B). Learners are provided with a scale and descriptions of what they can do based on their resulted level of proficiency. According to the CEFR, level A is defined as a basic user, level B as an independent user and level C as a proficient user. Once learners are provided with the description of their level, DIALANG also offers the option of *advisory* feedback which describes how they can improve their skills and proceed to the next level.

Research studies

Early studies have been conducted on the use of DIALANG. For example, a study conducted by Floropoulou (2002) investigated the attitudes of Greek and Chinese learners of English towards DIALANG and its self-assessment component. The study focused on ten subjects, five Greek and five Chinese students who were all inexperienced regarding self-assessment. The results showed that participants had a vague idea of what self-assessment really is and they tended to evaluate their language abilities unconsciously rather than consciously, depending on their cultural backgrounds. Regarding the effectiveness of DIALANG, Floropoulou mentions some technical issues that seem to have affected her findings, e.g. the cut-offs between the different levels, the factors that were taken into account in determining the score as well as the discrepancy between the self-assessment results and the test scores. As a result, it was suggested that test designers should manage the technical issues and make improvements in the system.

DIALANG feedback was investigated in the study conducted by Yang (2003). The study examined how twelve postgraduate students use the different types of feedback in the reading component and which factors influence their use through reading scores and student interviews. Findings revealed that elaborate feedback facilitates learning despite its diversity. Finally, the study also demonstrated the potential pedagogic advantage of computer-based tests.

The usefulness of the DIALANG language assessment system was also investigated by Huhta (2010) in his research study. 550 DIALANG users participated in his survey, and reacted positively to the system admitting that the feedback they received was useful. The system was also found to be useful for individual learners and their teachers, e.g. identifying the learners' level of proficiency, their strengths and weaknesses. The evidence collected indicated that many institutions use the system for placement purposes and that the most common problems concerned the vocabulary size placement test as well as some other technical aspects of the system.

Furthermore, Baglantzi (2012) explored the potential of DIALANG to serve placement purposes in the 1st grade of Greek state junior high school. The research focused on whether DIALANG can replace teacher-made placement tests for younger learners of English since placement tests are often criticized about their content, validity and reliability. For the purposes of the study, twenty students took the reading, writing and listening components of the DIALANG test and then reported on their experience through a questionnaire. Results showed that DIALANG can be used as a highly practical and useful placement tool for the skills tested. It was also found that the lack of time limit for the completion of DIALANG as well as the test level adaptability eliminated anxiety factors and increased the possibility of higher performance scores. However, findings demonstrated that both the teacher and the students lacked adequate familiarity with the DIALANG statements and the CEFR level descriptors, although the levels are used extensively in textbooks, language courses and exams in English Language Teaching (ELT) in Greece.

With regard to the practicality of the test, Baglantzi recommended the introduction of a specially-designed form where students could report the scores and feedback at the end of each subcomponent. Issues of practicality emerged during the administration of the listening test, where students were exposed to background noise which might have influenced their ability to perform well on the test. Moreover, it was recommended that DIALANG should be tailor-made for school placement purposes and be accessible from computer labs taken within school teaching periods of 45-60 minutes.

A more recent study conducted by Taghizadeh, Alavi and Rezaee (2014) focused on diagnosing 68 Iranian university students using DIALANG. The students were majoring in English language teaching (TEFL), literature (ELL) and translation (ELT). DIALANG self-assessment scales were used and results indicated that ELL students had the highest overall ranking especially in reading and writing. Lastly, one-way ANOVA between groups demonstrated statistically significant difference in the writing self-assessment statements for the three groups. The self-assessment statements of DIALANG were found to be effective in determining the level and language abilities of students. The researchers concluded that self-assessment should be promoted by language teachers to enable learners to reflect on their language proficiency level, take control of their learning process and formulate specific goals for their future progress.

From the limited but important research literature, it seems that DIALANG has been used to a large degree to examine concepts of self-assessment and feedback. Despite its technical deficiencies, most of the researchers have described DIALANG as useful, especially in terms of its extensive feedback and self-assessment scales. It has also been characterized as effective for fulfilling placement procedures and measuring the level and abilities of students. However, no study so far has focused on the possibility of using DIALANG to track the progress of students over time in academic contexts. Such research results could provide additional insights about the suitability of the test in such contexts and inform attempts to update and develop new and robust versions of DIALANG.

Methodology

Context of study and research questions

The Language Centre (LC) of the University of Cyprus (UCY) has been offering academic English (and other) language courses for many years. The overall goal of the English courses, in particular, is to enhance knowledge of the language and help students perform academically in a university context. According to the LC requirements, the expected level of incoming students is B1+ (Threshold) on the CEFR (Council of Europe, 2001) or approximately the level of Cambridge FCE Exam. As such, students are placed on the first level of their academic English language programme. After completing all three levels of their programme, students are expected to reach B2 level (https://www.ucy.ac.cy/langce/en/). In terms of monitoring the placement of the incoming students, the university system had very little in place. Therefore, in the absence of any placement procedures (e.g. a placement test), all students were registered in the same academic course, regardless of their language level. Also, in terms of recording students' language improvement, no evidence of the range or variability of students' English language levels within each cohort was systematically collected by the LC or the university departments that could be used as a yardstick for comparison. However, students' performance in the language courses was assessed every semester through summative practices, e.g. a mid-term and a final exam usually set by their course teachers. Depending on their final grade, students were able to move to the next level or repeat the same level course.

Given this current state of affairs, plus the need to moderate the quality of students' learning over time and evaluate the effectiveness of the academic English Language programme, DIALANG was employed by UCY as the most suitable test to diagnose students' progress in English due to its practicality, accessibility and immediate feedback. The choice was also made on the assumption that in terms of content, DIALANG and the LC courses would be compatible in that they were both linked to the same CEFR levels.

The study undertaken recorded the progress of three cohorts of university students (Department of Economics) after attending a three-level (semester long) compulsory English language course. As mentioned above, the aim of this study was to evaluate whether a diagnostic test such as DIALANG could be used to monitor the language development of university students over time, that is over the three semesters of English language classes offered at the LC of the current university. The research questions that this study attempted to answer were the following:

1) Is there any improvement in the scores of EFL university students over time when DIALANG is used? If yes, which skills show improvement?
2) Can DIALANG be used to measure language learners' progress over time?

Participants and administration

All student participants were in their first year of study when they took the DIALANG test for the first time. Their age varied between 18-24. Precise information on the incoming students' English language level of the was not available at the time of entry to the university. However, given the central role of language learning in the public and private sector in the country (see Lamprianou & Lamprianou, 2013; Tsagari, 2014; Xanthou & Pavlou, 2010; Lazarou et al. 2009), students had been exposed to a lot of opportunities for English language learning. Especially in the case of private language institutes, students attend English courses in order to improve their language skills and prepare for internationally recognized language certificate examinations (Tsagari, 2009; 2014) mainly at level B2 on the CEFR.

A total of 250 students across the three cohort years participated in the study. The first cohort was registered for English language courses in 2013, the second in 2014 and the third one in 2015. The cohorts of 2013 and 2014 took DIALANG three times: at the end of the first, second and third semester while the 2015 cohort took the test twice only, e.g. at the end of the first and second semester due to unforeseen circumstances. Table 1 below presents the overall number of participants per year.

Table 1. Number of participants for each cohort and gender

	Entry Year	Male (N)	Female (N)	Total (N)
Cohort 1	2013	34	64	**98**
Cohort 2	2014	29	53	**82**
Cohort 3	2015	32	38	**70**
Total		**95**	**155**	**250**

For each administration, student cohorts were usually divided into two smaller groups due to their size and lack of space in the university computer labs. Students were invigilated by an assistant who also helped them in recording their test results at the end of every subcomponent (as in Baglantzi, 2012). The administration of the DIALANG test per group lasted approximately three hours and took place during the regular exam period of the term. It requires less time (about 2-2.5 hours) when the test is taken individually, but time was extended because students were given a short introduction on the completion steps of each DIALANG subcomponent every time they took the test. This was deemed appropriate to avoid any misunderstandings that might have an impact on students' performance. Once students had completed a subcomponent of the test, they were asked to instantly save their scores in individual students' files that had been specially designed for that purpose and uploaded on their computers (Baglantzi, 2012). The procedure of score saving was crucial and students were reminded to do so before moving to the next subcomponent of the test.

All participants were asked to complete the subcomponents of VSPT, vocabulary, structures, reading and writing as well as the self-assessment test for reading and writing. The listening test was not taken due to lack of audio-visual equipment.

Data preparation

In order to prepare the data for statistical analysis, data cleaning was carried out. Missing values were identified due to the absence of some participants on the day of the test administration. Also, a number of students attained zero out of 1000 (which is the maximum score) in the VSPT. This was attributed mainly to the DIALANG reporting system, that is, the oversensitive (to guessing) algorithm of the VSPT which seems to have been the cause of so many zero points on the VSPT. To confirm this, in the analysis of the data, VSPT scores were analyzed twice, that is including and excluding zeros. In case of the latter, they were coded as missing values.

For statistical purposes, scores in the tests of vocabulary, structures, reading and writing were reported on the six levels of the CEFR (A1, A2; B1, B2; C1, C2). Scores were also substituted by numbers ranging from 1 to 6, where 1 describes the lowest level in the CEFR (e.g. A1) and 6 the highest (e.g. C2). Finally, since most cohorts completed the DIALANG test across three time points, the first time they took the test was coded 'Test 1', the second time was coded 'Test 2' and the third time 'Test 3' (see Table 2).

Statistical Analysis methods

Descriptive statistics

Stata was used in the current study for the analysis of data (StataCorp, 2013). Descriptive statistics for every cohort and time-points were calculated. This included the mean scores for each cohort and test. The mean scores for VSPT (including and excluding zeros) along with a percentage of the missing data were also included in the analysis. For each mean score, a 95% confidence interval was estimated. [2]

[2] A 95% confidence interval implies that we are 95% confident that the true mean scores in the general population will lie between a specific range. Wider confidence intervals represent higher levels of variability (Bland, 2015). The mean score and the 95% confidence interval for each subcomponent and the whole test, are displayed in the tables included in the results section.

Paired t-tests

In order to compare the performance of students across time, a two sample t-test was carried out[3]. This calculated the mean difference in scores between Test 1 and Test 2, Test 2 and Test 3 and Test 1 and Test 3. The expectation here was that there would be an improvement over time, e.g. between the first time they took the test (Test 1) and the third time (Test 3) and, therefore, the means would be different. The VSPT scores were also added to the paired t-test analysis to examine if the improvement in the baseline VSPT scores is consistent with the improvement of other subcomponents.

ANOVA

In order to compare the performance of students across time in different cohorts, an ANOVA (ANalysis Of VAriance) test was carried out to compare the mean scores between the three cohorts for each test (Test 1, Test 2, Test 3) and thus to determine the cohort with the highest performance scores. As with the t-test, the level of significance was set to 0.05 (Bland, 2015).

Finally, to find out which cohort means were different, two-sample t-tests were conducted to test for any difference in means of each pair of student cohorts. This procedure is prone to multiple hypotheses testing which can lead to biased results (Field, 2009). To adjust any bias in the post-ANOVA t-tests, we used Tukey's test which changes the significance level to adjust for performing multiple hypotheses (see Appendix E).

Longitudinal analysis

For a more robust analysis of the data, longitudinal analysis of panel data using a generalized linear model was carried out. This was used to provide predictions for the mean improvements of each cohort, adjusting for both cohort and time-point. Significance testing was performed for each model coefficient and its associated test statistic (z-score) and *p* value. These are also reported (see Appendix G).

Sensitivity analysis

As part of a sensitivity analysis, mean scores for each subcomponent and test were calculated only for those who took all three tests. Additionally, the students with no missing test scores were included in a separate longitudinal regression. This was carried out to ensure that there was no difference in the mean improvement between the overall sample and those who took the test all three times (see Appendix I).

[3] The paired t-test calculates the difference within each before-and-after pair of measurements, estimates the mean of these changes, and tests whether this mean of the differences is statistically significant (Field, 2009).

Results

Overall improvement

Table 2 presents the overall number of students who took the DIALANG test across the three time points. Although the total number of participants was 250, at any one time, the number of students with valid data was smaller due to non-responses (Table 2).

Table 2. Mean scores across tests

	Test 1				Test 2				Test 3			
Subcomponent	**N**	**Missing (%)**	**Mean**	**(95% CI)**	**N**	**Missing (%)**	**Mean**	**(95% CI)**	**N**	**Missing (%)**	**Mean**	**(95% CI)**
Vocabulary	230	8.0	3.26	(3.14, 3.38)	161	35.6	3.39	(3.25, 3.52)	143	42.8	3.50	(3.35, 3.64)
Structures	228	8.8	3.22	(3.08, 3.36)	161	35.6	3.25	(3.09, 3.41)	143	42.8	3.26	(3.09, 3.42)
Reading	229	8.4	2.40	(2.25, 2.55)	161	35.6	2.39	(2.20, 2.58)	143	42.8	2.40	(2.21, 2.59)
Writing	230	8.0	2.40	(2.28, 2.52)	161	35.6	2.29	(2.13, 2.45)	142	43.2	2.46	(2.31, 2.62)
Average	228	8.8	2.81	(2.70, 2.93)	161	35.6	2.83	(2.69, 2.97)	142	43.2	2.90	(2.76, 3.04)
VSPT												
including 0	228	8.8	302.0	(265.1, 339.0)	161	35.6	322.9	(280.6, 365.3)	143	42.8	278.2	(232.9, 323.5)
excluding 0	169	32.4	407.5	(368.8, 446.2)	124	50.4	419.3	(377.4, 461.2)	98	60.8	405.9	(357.9, 454.0)

Overall, the results showed that students scored higher in the subcomponents of vocabulary and structures across the three test administrations (see Table 2). The CEFR level corresponding to the vocabulary and structures test is estimated between B1-B2 (generated by DIALANG). In the subcomponents of reading and writing, students performed lower with a corresponding level of A2-B1. Similar results are presented in Table 3.

In terms of overall improvement from Test 1 to Test 3, students did better in the subcomponents of vocabulary, structures and writing. In the subcomponent of reading, the mean scores remained relatively stable. The average mean scores indicate that students performed better in the third test administration; however, their scores did not exceed level A2-B1 on the CEFR.

The average mean scores of VSPT including zeros were between 201-400 while those excluding zeros were between 401-600. According to the placement test feedback (see Appendix C), users who score at the level of 201-400, "have a limited vocabulary which may be sufficient for ordinary day-to-day purposes, but probably doesn't extend to more specialist knowledge of the language" (https://www.lancaster.ac.uk/researchenterprise/dialang/about). Similarly, users who score 401-600 are described as having good vocabulary but may be facing difficulties "handling material that is intended for native speakers" (see Appendix C). In general, VSPT scores increased in the second test administration and then decreased in the third time point. As it was expected, excluding zeros (counting them as missing) generated VSPT scores that were, on average, 30-37% higher than those generated when including zeros.

To further address the research questions, paired t-test analysis on the 'before and after' comparisons was undertaken based on data from students who took the tests (see Table 3).

Table 3. Paired t-test for each test comparison

Test 1 vs Test 2					
Subcomponent	**N**	**Mean difference**	**(95% CI) ***	**t statistic**	***p* value**
Vocabulary	156	0.12	(0.03, 0.20)	2.71	0.008
Structures	156	0.04	(-0.08, 0.16)	0.62	0.538
Reading	156	0.05	(-0.07, 0.18)	0.82	0.416
Writing	156	-0.11	(-0.24, 0.02)	-1.70	0.091
Average	156	0.02	(-0.05, 0.09)	0.67	0.501
VSPT (including 0)	154	24.47	(-9.58, 58.51)	1.41	0.158
VSPT (excluding 0)	99	35.05	(-8.53, 78.63)	1.60	0.114
Test 2 vs Test 3					
	N	**Mean difference**	**(95% CI)**	**t statistic**	***p* value**
Vocabulary	90	0.08	(-0.05, 0.20)	1.22	0.225
Structures	90	-0.01	(-0.15, 0.13)	-0.16	0.877
Reading	90	-0.02	(-0.17,0.13)	-0.29	0.770
Writing	90	0.21	(0.06, 0.37)	2.70	0.008
Average	90	0.06	(-0.02, 0.15)	1.53	0.129
VSPT (including 0)	90	-47.53	(-89.02, -6.04)	-2.28	0.025
VSPT (excluding 0)	48	-52.88	(-113.85, 8.10)	-1.74	0.088
Test 1 vs Test 3					
	N	**Mean difference**	**(95% CI)**	**t statistic**	***p* value**
Vocabulary	138	0.30	(0.20, 0.41)	5.63	<0.001
Structures	137	0.12	(0.004, 0.24)	2.04	0.043
Reading	137	0.07	(-0.08, 0.21)	0.88	0.382
Writing	138	0.18	(0.06, 0.31)	2.89	0.005
Average	137	0.17	(0.10, 0.24)	4.97	**<0.001**
VSPT (including 0)	136	23.96	(-14.18, 62.11)	1.24	0.216
VSPT (excluding 0)	75	41.09	(-10.13, 92.31)	1.60	0.114

*CI= Confidence Interval

According to the findings, a significant improvement can be seen in the subcomponents of vocabulary, structures and writing in at least one test comparison. For vocabulary, there was a significant improvement from Test 1 to Test 2 (p=0.008) and from Test 1 to Test 3 (p< 0.001). For structures there was a significant improvement only from Test 1 to Test 3 (p=0.043). The subcomponent of reading did not demonstrate improvement in any of the tests (p>0.05). For the first and second tests, the writing scores were relatively similar, while the difference in scores between Test 1 and Test 3 as well as Test 2 and Test 3 were statistically significant (p=0.005 and p=0.008 respectively). Students showed a statistically significant difference between the average mean scores of Test 1 and Test 3, implying an overall improvement (mean difference=0.17, p=<0.001). As mentioned above, students performed better in the subcomponents of vocabulary and structures with a corresponding CEFR level of B1-B2. They performed lower in the skills of reading and writing, attaining a level between A2-B1 on the CEFR scale.

Improvement in VSPT scores was not significant for any time point as the *p* values were all >0.05. Although including zero VSPT scores resulted in a substantially lower mean score in any given test, this had no impact on the difference between any two tests measured using paired t-tests (Table 3). For instance, on average, students scored 24.5 points higher in the second test than the first test in the method including zeros. When excluding zeros, students scored about 35 points higher in the second test compared to the one. Although this is slightly higher, it is not significantly higher considering the placement test is scored out of 1000. Therefore, the way that zeros are dealt with in the VSPT improvement analysis does not affect the conclusions drawn.

The results presented in Tables 2 and 3 indicate that there are discrepancies between the expected level upon completion of the courses (B2 level according to the LC) and the level obtained in the DIALANG test. Particularly, with regard to reading and writing, the incoming level of students was lower than B1+ (required by the LC) and the level after the completion of the three levels of English did not exceed B1 (B2 level is required by the LC). The same applies to the skills of vocabulary and structures where although students achieved a higher mean score, the level remained between B1-B2, even after attending three semesters of academic English language courses. The average test results (Table 2) demonstrate that students performed at A2-B1 level on the CEFR when starting the compulsory language courses and remained at that level even after completing three semesters. These results raise questions about possible reasons that might have been the source of such low scores such as the effectiveness of the language courses and other factors. These will be discussed in more detail in later sections.

Cohort effects

Appendix D reports the mean scores for every cohort across the three test administrations. The mean scores for each skill suggest that Cohort 3 performed consistently better in all skills in Tests 1 and Tests 2, while Cohort 1 performed better in Test 3, particularly in the skills of vocabulary, reading and writing. The average mean scores demonstrate that Cohort 3 attained higher scores in the first two tests (Test 1 M=3.06, Test 2 M=3.01, see Appendix D) while Cohort 2 performed consistently lower.

Apart from the summary presented in Appendix D, the full cohort comparison output of the ANOVA tests is shown in Appendix E. For vocabulary, Cohort 3 had a significantly higher mean score than Cohort 2 for Test 1 (F=3.18, p=0.043, see Appendix E). This was also true for writing (F=4.57, p=0.011). For all other comparisons, although Cohort 3 performed better than the rest, there was no strong evidence for a difference between the cohorts (ANOVA test: p>0.05).

Improvements in the scores by cohort for each pair of tests, along with the paired t-test results, are reported in Appendix F. There was no significant difference in the mean scores for reading between any two cohorts for any of the tests (paired t test: p>0.05). For vocabulary, students performed better in Tests 2 and 3 and this was true only for Cohort 1 and 2 (see Appendix F for *t* statistics and *p* values). The scores of Cohort 2 and 3 in structures did not increase over time but Cohort 1 showed a significant difference in the mean scores for structures between Test 1 and Test 3. Finally, Cohort 1 improved in writing but Cohort 3 writing scores decreased significantly in Test 2. Although Cohort 3 obtained the highest scores in Test 1 and Test 2 (Appendix D), there was no significant progress in any of the skills[4] (Appendix F). A statistically significant difference can be observed in the average mean scores of Test 1 vs Test 2 and Test 1 vs Test 3 for Cohort 1. This suggests an overall improvement from the first test administration to the third one (p=0.035 and p<0.001 respectively).

Similar results have been generated by analyzing the data over time. Appendix G shows that there was a significant improvement over time in Cohort 1 for most subcomponents, including vocabulary, structures and writing. Cohort 2 improved significantly only in the skill of vocabulary while the performance of Cohort 3 in writing decreased significantly.

The corresponding CEFR level attained by Cohort 1 and Cohort 2 was between A2-B1 across all three test administrations (see Appendix D). The results from Cohort 3 correspond to B1 level in Test 1 and Test 2.

[4] However, this should be interpreted with caution as Cohort 3 did not take DIALANG for a third time.

Appendix H depicts all possible trajectories of CEFR levels across time. This gives a more detailed picture of the progress in CEFR level for each individual student for a given skill.

The results of the sensitivity analysis are shown in Appendix I. There were no significant differences between the overall sample and those who took the test all three times.

Discussion of results

Summary of the main findings

Although expected to enter the university with a minimal B1+level on the CEFR level and attain B2 at the end of three semesters of compulsory English study, the reality for the students who participated in the current study was different. For example, even though findings revealed an overall improvement in the scores of students across the three test administrations, this improvement was not significant in terms of the CEFR, since students attained only level A2-B1 after their three semesters of English language tuition. In terms of the various DIALANG subcomponents, students demonstrated overall improvement over time in vocabulary, structures and writing but no significant improvement in reading. More specifically, students scored a corresponding level of B1-B2 in vocabulary and structures and a corresponding level of A2-B1 for reading and writing.

The results also showed cohort differences. Significant improvement over time was achieved only by Cohort 1. However, Cohort 3 seemed to have performed consistently better while Cohort 2 attained consistently lower scores. In terms of the CEFR scales, none of the cohorts exceeded B1 level, even after attending all three levels of the English language programmes offered at the current university.

Discrepancy between the expected level and the observed level

All together, the results point to discrepancies between the expected level and the observed level of the students. According to the LC, students were expected to perform at a B2 level after completing three semesters of English language training. However, the findings demonstrated that students did not reach the expected level of B2 but rather remained at A2-B1 level. This result is somewhat disappointing as it seems that, despite their three semesters of English language courses, students' level has not improved. However, before drawing definite conclusions, these need to be further explored through additional measures, e.g. the examination of the teaching content of the courses offered or examination of the compatibility of DIALANG to the course objectives and course content offered. These and other measures could have motivated appropriate alignment and modifications in the English course syllabi and/or teaching at initial stages of the study to ensure that students reach the desired CEFR level.

Focus on skills

Findings also suggested that students faced difficulties in the skills of reading and writing compared to the skills of vocabulary and grammar (structures). However, it is difficult to draw any definite conclusions since there was no information available to the researchers of whether the LC's teaching practices favoured certain skills over others. Also, while writing is tested indirectly in DIALANG, e.g. users produce a limited amount of words, in the usual practice of writing assessment in the LC academic course, students produced extensive pieces of writing. Therefore, it is not possible to say with any degree of certainty that the results of the writing component in DIALANG could be attributed to the quality of the LC programmes only. Perhaps other than the careful examination of the course syllabus to determine any particular focus on specific skills, semi-structured interviews with tutors of the LC and their students could have been conducted. These would determine the focus on any particular skills, teaching practices or reasons behind missing items (and values) in the current DIALANG administrations.

Cohort differences

Cohort discrepancies is another finding that merits further discussion. For example, unlike Cohort 1 and 2, Cohort 3 exhibited homogeneity in scores reflected in the almost equal level of attainment during the first and second test administrations (Appendix D). One explanation could be that the organization and administration of the DIALANG test for Cohort 3 was more effective given the experience accumulated after the administrations of Cohort 1 and 2 (also reflected in less missing values). For these reasons, perhaps Cohort 3 managed to achieve better results and uniformity in scores as opposed to the other two cohorts.

Challenges

As with every type of research, conducting a large-scale study has had particular challenges. One of the major challenges of this study was the number of missing data. Students were not consistently present in all three administrations of DIALANG. The number of participants constitutes a major concern for most research studies since participants can choose to withdraw from the study at any time given (Croninger & Douglas, 2005). The same limitation applied to Cohort 3 which, due to unforeseen circumstances, did not provide scores for the third test administration and, therefore, an accurate comparison between the three time-points was not possible.

Additionally, even during the same administration of the test, further missing data was detected in the subcomponents. This could be due to the fact that students either skipped one of the subcomponents, or moved on to the next one without saving their results. As students were taking a computer-based test, it was difficult, given the limited sources available, to monitor the completion of all the subcomponents.

Associated with missing data is the way it can affect the external validity of the study and raise questions over whether the sample used in the analysis is a truly representative sample of the general population (Kadam & Bhalerao, 2010). Estimates produced from the statistical analysis are less biased if the data is missing at random, which is what was also assumed in the current study. However, we do not know if those students who did not take DIALANG a second or third time, or those who did not agree to participate from the beginning of the study, would have scored higher or lower in DIALANG. Therefore, the mean scores reported here may have been overestimated.

Another challenge involved the VSPT and issues of reliability. Reliability is an essential quality when it comes to determining the effectiveness and usefulness of a language test (Bachman & Palmer, 1996). Huhta (2010) explains that in his DIALANG study some test takers received low scores, which were incompatible with their level of proficiency, or even zero points. In this study, a number of students received zero scores in their VSPT tests. Their zero scores may have also affected the scores that students obtained in subsequent subcomponents, since the VSPT is used to place the learners at the appropriate level. This has been identified as a limitation of the DIALANG testing system and further development needs to be carried out so as to tackle any issues that may affect the accurate representation of students' abilities.

Further limitations, include aspects of the study design. DIALANG's scoring system which is based on the CEFR scales, may not have been particularly suitable for purposes of monitoring progress, since the scales are not well-refined and contain a small number of possible scores. For this reason, perhaps numerical answers of the items that the learners got correct would reflect more accurately the learners' proficiency. Also, the limited number of DIALANG test items should also be taken into account in further revisions of DIALANG especially in the case of writing.

Students' familiarity with test items could have also influenced the test scores. In other words, the attainment of higher scores in the third test administration may also be attributed to familiarity with items instead of true learning. Although taking the same test over a period of time provides an accurate measurement for purposes of comparison, higher scores may have been the result of becoming familiar with or even remembering the answer to certain questions and not a matter of improvement.

Likewise, comparing DIALANG scores with other measures of student achievement, before, during, and at the end of their programme, would have been useful in determining the validity of the DIALANG system in monitoring progress over time. Obtaining additional information about students' proficiency, e.g. other placement or summative test results as part of students' placement or in-course assessment of performance on units of study or end of the semester, would have enabled comparisons with DIALANG results and led to discussions of its effectiveness in the current context.

Suitability of DIALANG for tracking progress over time

A test developed for a particular purpose is not automatically valid for other purposes. DIALANG underwent piloting with a large number of learners and it has been used by various institutions for placement purposes. It has also been used in research as a measure of learners' language proficiency in reading (Alderson, Huhta, & Nieminen, 2016). This indicates that when it comes to measuring learners' English skills expressed in terms of CEFR levels, DIALANG is as valid as any other test based on general language rather than language for specific purposes. LSP (Language Specific Purpose) tests are defined as 'those involving language for academic purposes and for occupational or professional purposes' (Douglas, 2000, p.2). Nevertheless, DIALANG may not be suitable for the purpose of tracking development of students' language skills in a university setting for two main reasons. Firstly, DIALANG tests general English and not academic English. Academic courses focus on domain specific language skills, the development of which cannot be captured by a general language test. Students' general language skills probably develop too, but the discrepancy between the content of the academic courses and the test may explain why there was such little progress from the first to the third time that students took the test. Secondly, the CEFR levels that DIALANG uses as the test score are very wide. Therefore, the language programme must be sufficiently long and intensive in order to develop learners' proficiency to a degree that can be detected by such a relatively insensitive instrument such as DIALANG. These issues, combined with learners' performance and the test behaviour provide some answers to the question of the suitability of DIALANG for the purposes of tracking students' language progress over time.

Some more issues concerning the suitability of the test were its length and the recording of scores. The test lasted approximately three hours and was administered at the end of the semester when students were already in their exam period. This factor, might have also affected the reliability of scores, since students might have been influenced by exam fatigue and anxiety. The interface of DIALANG, which is rather outdated and not particularly user-friendly, might have also impacted on the test results.

Furthermore, the fact that students were requested to copy/paste their results in specially-designed folders on their computers after the administration of each subcomponent might have had an influence on students' performance as transfer time might have delayed the test taking process introducing extra fatigue. Even though for purposes of validity and reliability of the study the process of taking DIALANG was explained in detail before every test administration, this took a lot more time and along with saving the students' scores at the end of every subcomponent it turned out to be more time-consuming than expected requiring the employment of more than one assistant. A mechanism for score saving would add to the effectiveness of the test and eliminate time and other practical problems.

Despite the above challenges, DIALANG provides useful diagnostic information to its users. In the event DIALANG is used with large groups of students in the future, effective time sources and available spacious labs need to be considered. Assessing all skills may not be feasible when targeting large groups due to time issues, but teachers/administrators may focus on assessing fewer skills according to the needs of their students and their assessment purposes.

Conclusion

Despite the limitations of the current study, the findings can serve as a basis for conducting future research on the use of DIALANG with large groups of students. Specifically, using DIALANG to measure the language proficiency of students from other departments can yield substantial evidence for establishing the reliability of the DIALANG test. This can be done at the beginning of the semester and prior to the enrollment of students in academic English language courses. Such practice can produce significant results as to the level and needs of the incoming English (and other) language students. These future research endeavours should be pursued on the condition that a number of factors (e.g. purpose and context of testing), should not differ from those for which DIALANG is designed. Future studies where test conditions need to be strictly controlled to ensure comparable results among individuals within cohorts, at different times, and across cohorts can yield rich results. Also, participation levels of students need to be monitored very carefully to avoid fluctuation at different times and across subcomponents of the test. Nevertheless, the current research, despite the challenges it faced, is interesting in that it investigates the use of DIALANG over time in a learning environment which is in itself challenging; real-world research is not always seamless but can offer practical solutions and suggestions for future research.

In conclusion, DIALANG has been a pioneering example of innovative diagnostic computer-delivered, open-access second and foreign language assessment since the late 1990s (Alderson, 2005). For DIALANG to remain relevant it should take into consideration the suggestions made in this paper as well as align with state-of-the-art conceptualisations of diagnosis (e.g. Alderson et al., 2015a; Harding et al., 2015). The existing version of DIALANG represents a general approach to diagnostic language assessment; however, newer conceptualisations of diagnosis need to be considered (Alderson, et al., 2015a, b; Harding, et al., 2015) along with new and considerably modified approaches to computer-based diagnostic language assessment (Tsagari & Banerjee, 2015). Finally, plans to expand the assessment of writing, and to introduce diagnostic assessment of speaking (also through self- and peer-assessment), and update the automated diagnostic and recording scoring systems should be on the agenda of future revisions of DIALANG.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015a). Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields. *Applied Linguistics*, *36*(2), 236–260.

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015b). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320.

Alderson, J. C., Huhta, A., & Nieminen, L. (2016). Characteristics of Weak and Strong Readers in a Foreign Language. *The Modern Language Journal, 100*(4), 853-879.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baglantzi, V. (2012). Online Diagnostic Assessment: Potential and Limitations (The case of DIALANG in the Greek Junior High School Context). *Research Papers in Language Teaching and Learning*, *3*(1), 293-310.

Bland, M. (2015). *An Introduction to Medical Statistics* (4th ed.). Oxford: Oxford University Press.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Education.

Council of Europe. (2001). *A Common European Framework of Reference for Languages. Learning, Teaching, Assessment.* Strasbourg: Council of Europe.

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research. *New Directions for Institutional Research*, 127: 33–49.

Davies, A. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

DIALANG. https://dialangweb.lancaster.ac.uk/

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge University Press.

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London: SAGE Publications.

Floropoulou, C. (2002). *Foreign language learners' attitudes to self-assessment and DIALANG: a comparison between Greek and Chinese learners of English*. (Unpublished Master's thesis). Lancaster University, Lancaster, England.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123-144.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336.

Harding, L., Brunfaut, T., Huhta, A., Alderson, J. C. & Fish, A. (2018) *DIALANG 2.0: Charting a course for revision and expansion of an online diagnostic testing system*. Paper presented at 15th EALTA Conference 'Technology-Based Language Assessment: Benefits and Challenges', 25 – 27 May 2018, Bochum, Germany.

Huhta, A. (2010). *Innovations in diagnostic assessment and feedback: An analysis of the usefulness of the DIALANG language assessment system*. (Unpublished Doctoral Dissertation). University of Jyväskylä, Jyväskylä, Finland.

Huhta, A., & Figueras, N. (2004). Using the CEF to promote language learning through diagnostic testing. In K. Morrow (Ed.), *Insights from the Common European Framework* (pp. 65–76). Oxford: Oxford University Press.

Kadam, P., & Bhalerao, S. (2010). Sample size calculation. *International Journal of Ayurveda Research*, *1*(1), 55–57.

Klimova, B. F., & Hubackova, S. (2013). Diagnosing Students' Language Knowledge and Skills. *Procedia-Social and Behavioral Sciences*, *82*, 436–439.

Lamprianou, I., & Afantiti Lamprianou, T. (2013). Charting Private Tutoring in Cyprus: A Socio-Demographic Perspective. In Bray, E. M., Mazawi, A., & G. R. Sultana (Eds.), *Private Tutoring across the Mediterranean: Power Dynamics and Implications for Learning and Equity* (pp.29-56). Rotterdam: SensePublishers.

Lazarou, C., Panagiotakos, B. D., Kouta, C., & Matalas, L. A. (2009). Dietary and other lifestyle characteristics of Cypriot children: results from the nationwide CYKIDS study. *BMC Public Health*, *9*(1):147.

Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing* (3rd ed.). Taipei: Tung Hua Book Company.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal*, *76*(4), 513–521.

StataCorp, L. (2013). Stata: release 13-statistical software. *College Station, TX*.

Taghizadeh, M., Alavi, S. M., & Rezaee, A. A. (2014). Diagnosing L2 learners' language skills based on the use of a web-based assessment tool called DIALANG. *International Journal of E-Learning & Distance Education*, *29*(2), 1–28.

Tsagari, D., & Banerjee, J. (2015). Language Assessment in the Educational Context. In M. Bigelow & J. Ennser-Kananen (Eds.), *Handbook of Educational Linguistics* (pp. 339–352). New York: Routledge/Taylor & Francis Group.

Tsagari, D. (2014). Investigating the face validity of Cambridge English First in the Cypriot context. *Research Notes*, *57*(1), 23–31.

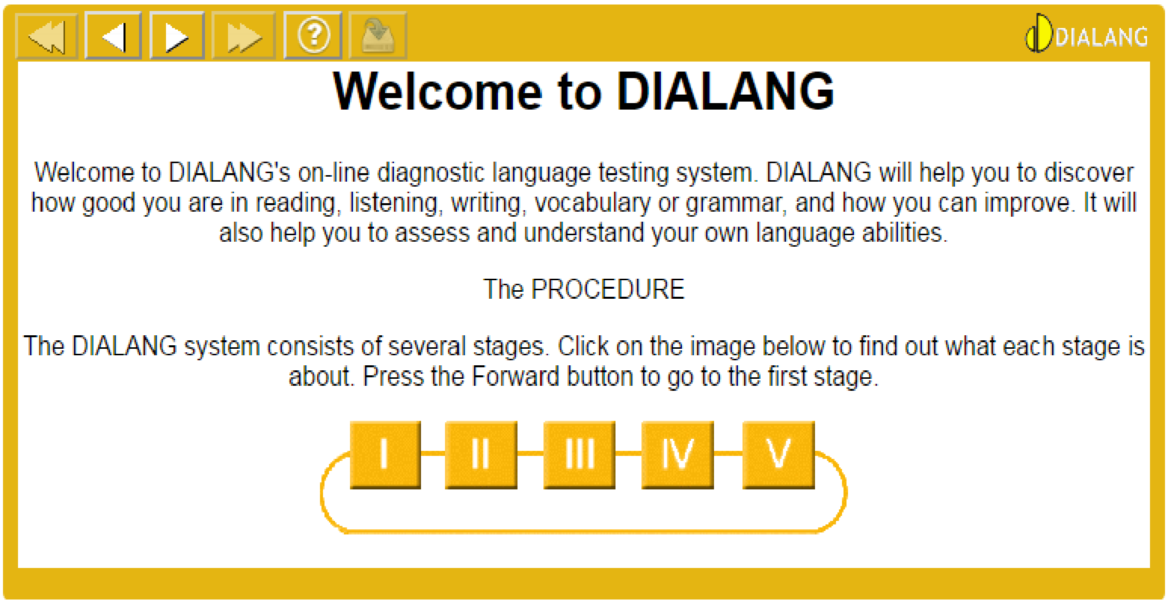
Tsagari, D. (2009). *The Complexity of Test Washback: An Empirical Study*. Frankfurt am Main: Peter Lang.

Yang, R. (2003). Investigating how test-takers use the DIALANG feedback. (Unpublished Master's thesis) Lancaster University, Lancaster, UK.

Xanthou, M., & Pavlou, P. (2010). Teachers' perceptions of students' attitudes in mixed ability EFL state primary school classes. In Psaltou-Joycey, A., & M. Mattheoudakis (Eds.), *Advances in Research on Language Acquisition and Teaching – Selected Papers. Proceedings of Greek Applied Linguistics Association 14th International Conference* (pp. 473-485). Thessaloniki: GALA.

Appendices

Appendix A: The DIALANG test

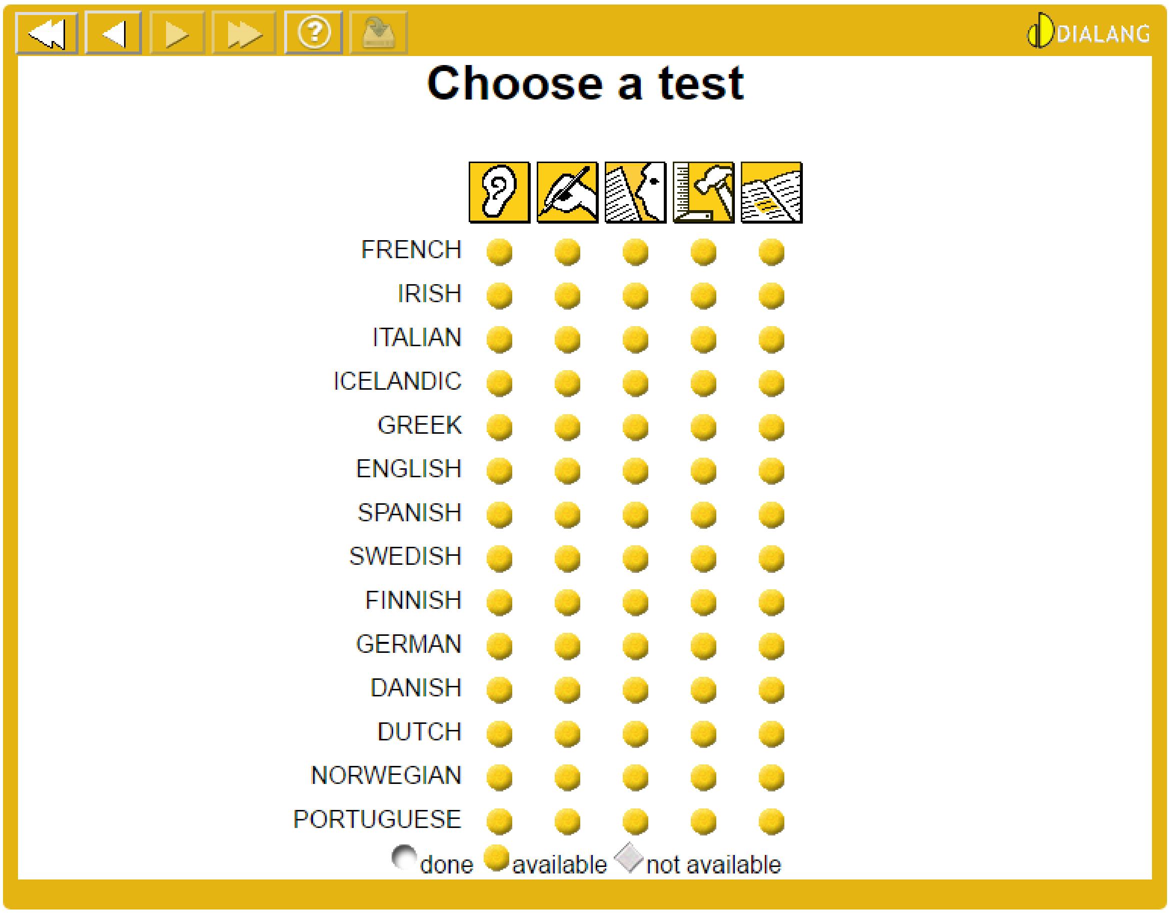


Welcome to DIALANG

Welcome to DIALANG's on-line diagnostic language testing system. DIALANG will help you to discover how good you are in reading, listening, writing, vocabulary or grammar, and how you can improve. It will also help you to assess and understand your own language abilities.

The PROCEDURE

The DIALANG system consists of several stages. Click on the image below to find out what each stage is about. Press the Forward button to go to the first stage.

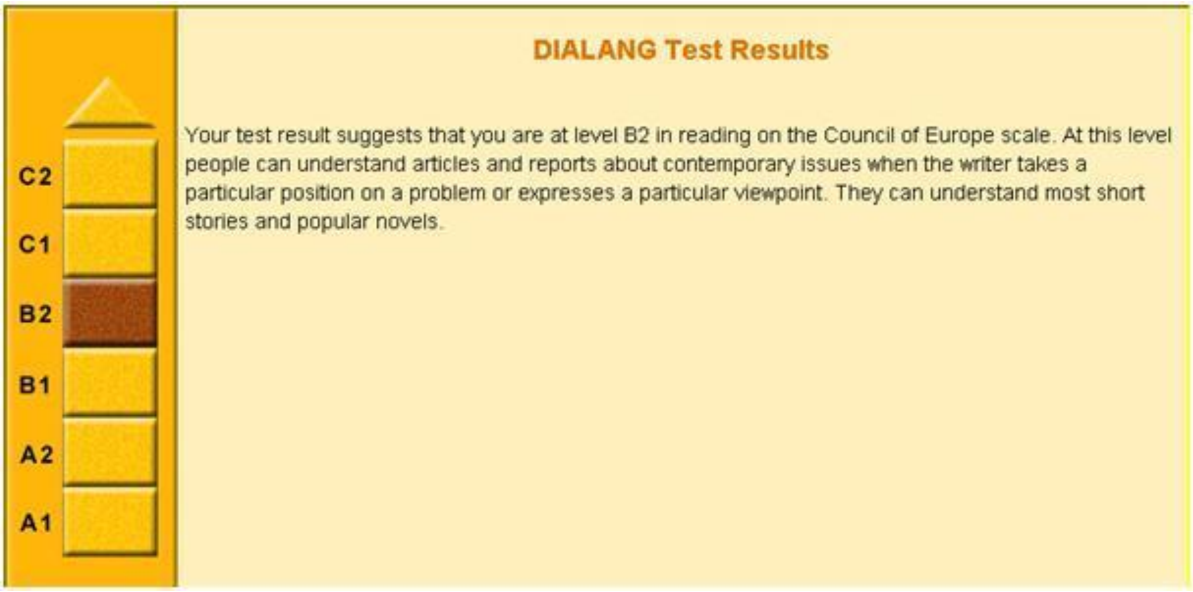


Choose a test

FRENCH	●	●	●	●	●
IRISH	●	●	●	●	●
ITALIAN	●	●	●	●	●
ICELANDIC	●	●	●	●	●
GREEK	●	●	●	●	●
ENGLISH	●	●	●	●	●
SPANISH	●	●	●	●	●
SWEDISH	●	●	●	●	●
FINNISH	●	●	●	●	●
GERMAN	●	●	●	●	●
DANISH	●	●	●	●	●
DUTCH	●	●	●	●	●
NORWEGIAN	●	●	●	●	●
PORTUGUESE	●	●	●	●	●

○ done ● available ◇ not available

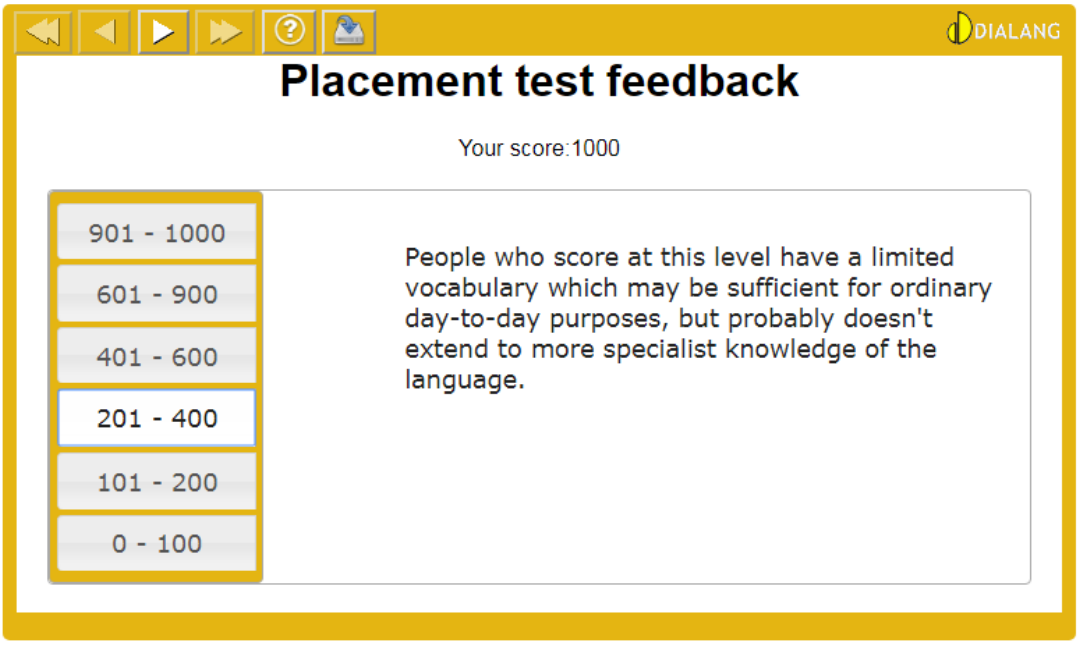
Appendix B: Your level



DIALANG Test Results

Your test result suggests that you are at level B2 in reading on the Council of Europe scale. At this level people can understand articles and reports about contemporary issues when the writer takes a particular position on a problem or expresses a particular viewpoint. They can understand most short stories and popular novels.

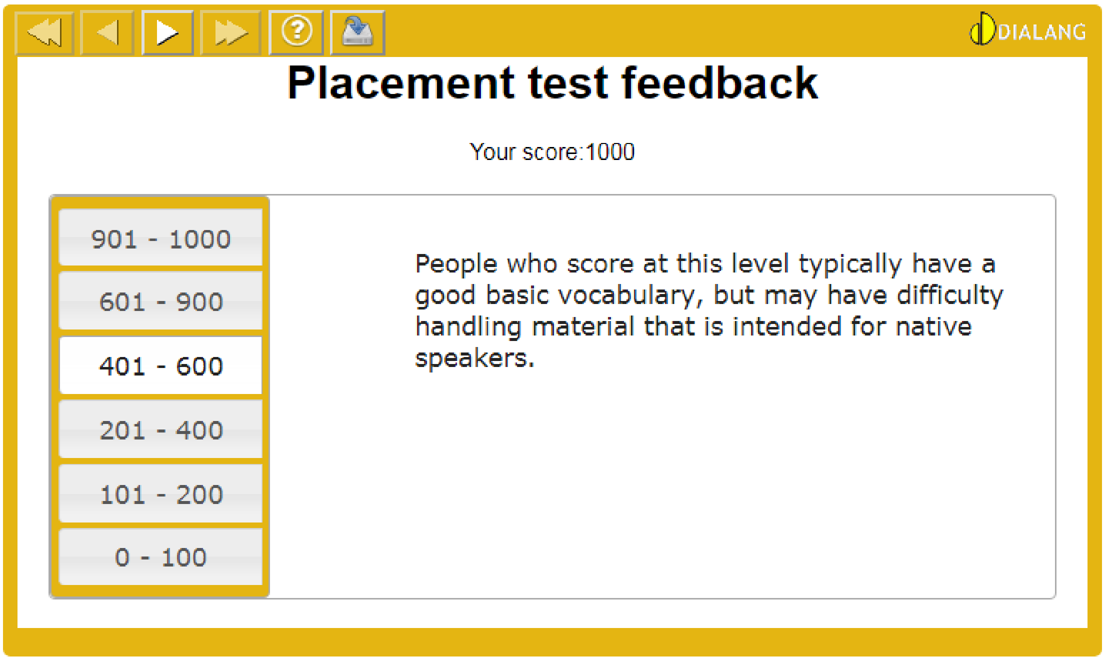
Appendix C: Vocabulary Size Placement Test



Placement test feedback

Your score:1000

People who score at this level have a limited vocabulary which may be sufficient for ordinary day-to-day purposes, but probably doesn't extend to more specialist knowledge of the language.



Placement test feedback

Your score:1000

People who score at this level typically have a good basic vocabulary, but may have difficulty handling material that is intended for native speakers.

Appendix D

Number of students (N) and mean score by cohort

		Test 1			Test 2			Test 3		
Cohort	**Subcomponent**	**N**	**Mean**	**(95% CI)**	**N**	**Mean**	**(95% CI)**	**N**	**Mean**	**(95% CI)**
Cohort 1	**Vocabulary**	95	3.24	(3.04, 3.44)	38	3.34	(3.05, 3.63)	80	3.55	(3.37, 3.73)
	Structures	93	3.22	(2.99, 3.44)	38	3.16	(2.80, 3.52)	80	3.35	(3.12, 3.58)
	Reading	94	2.48	(2.26, 2.70)	38	2.37	(1.93, 2.81)	80	2.55	(2.30, 2.80)
	Writing	95	2.36	(2.17, 2.54)	38	2.32	(2.01, 2.62)	80	2.54	(2.34, 2.74)
	Average	93	2.81	(2.62, 2.99)	38	2.80	(2.48, 3.11)	80	3.00	(2.81, 3.19)
Cohort 2	**Vocabulary**	75	3.09	(2.88, 3.30)	61	3.28	(3.07, 3.48)	63	3.43	(3.20, 3.66)
	Structures	75	3.04	(2.78, 3.30)	61	3.10	(2.87, 3.32)	63	3.14	(2.92, 3.37)
	Reading	75	2.15	(1.90, 2.39)	61	2.21	(1.96, 2.47)	63	2.21	(1.93, 2.50)
	Writing	75	2.21	(2.01, 2.41)	61	2.11	(1.88, 2.35)	62	2.37	(2.13, 2.61)
	Average	75	2.62	(2.42, 2.82)	61	2.68	(2.48, 2.88)	62	2.78	(2.57, 2.98)
Cohort 3	**Vocabulary**	60	3.50	(3.28, 3.71)	62	3.53	(3.31, 3.75)	-	-	-
	Structures	60	3.45	(3.20, 3.70)	62	3.47	(3.21, 3.73)	-	-	-
	Reading	60	2.60	(2.28, 2.92)	62	2.58	(2.27, 2.89)	-	-	-
	Writing	60	2.70	(2.43, 2.97)	62	2.45	(2.17, 2.73)	-	-	-
	Average	60	3.06	(2.83, 3.30)	62	3.01	(2.78, 3.24)	-	-	-

Appendix E

ANOVA tests for cohort comparisons

		Test 1				Test 2				Test 3			
Subcomponent	**Cohort comparison**	**Mean difference**	**(95% CI)**	**Tukey's t**	***p* value**	**Mean difference**	**(95% CI)**	**Tukey's t**	***p* value**	**Mean difference**	**(95% CI)**	**Tukey's t**	***p* value**
Vocabulary	**Cohort 1 vs 2**	-0.15	(-0.48, 0.19)	-1.03	0.559	-0.06	(-0.49, 0.36)	-0.35	0.933	-0.12	(-0.41, 0.17)	-0.83	0.405
	Cohort 2 vs 3	0.41	(0.02, 0.80)	2.51	0.034	0.25	(-0.12, 0.62)	1.63	0.238	-	-	-	-
	Cohort 1 vs 3	0.26	(-0.11, 0.62)	1.67	0.219	0.19	(-0.23, 0.61)	1.97	0.536	-	-	-	-
		Anova test: F=3.18, p=0.043				Anova test: F=1.40, p=0.249				Anova test: F=0.70, p=0.405			
Structures	**Cohort 1 vs 2**	-0.18	(-0.57, 0.22)	-1.04	0.551	-0.06	(-0.55, 0.43)	-0.29	0.956	-0.21	(-0.54, 0.12)	-1.24	0.216
	Cohort 2 vs 3	0.41	(-0.03, 0.85)	2.19	0.075	0.37	(-0.06, 0.80)	2.04	0.107	-	-	-	-
	Cohort 1 vs 3	0.23	(-0.19, 0.66)	1.31	0.390	0.31	(-0.18, 0.80)	1.50	0.296	-	-	-	-
		Anova test: F=2.39, P=0.094				Anova test: F=2.30, p=0.103				Anova test: F=1.55, p=0.216			
Reading	**Cohort 1 vs 2**	-0.33	(-0.75, 0.08)	-1.89	0.144	-0.16	(-0.74, 0.43)	-0.63	0.805	-0.34	(-0.72, 0.03)	-1.82	0.072
	Cohort 2 vs 3	0.45	(-0.01, 0.92)	2.31	0.057	0.37	(-0.14, 0.88)	1.70	0.666	-	-	-	-
	Cohort 1 vs 3	0.12	(-0.32, 0.56)	0.65	0.795	0.21	(-0.37, 0.79)	0.86	0.207	-	-	-	-
		Anova test: F=3.02, p=0.051				Anova test: F=1.46, p=0.235				Anova test: F=3.29, p=0.072			
Writing	**Cohort 1 vs 2**	-0.14	(-0.49, 0.20)	-0.99	0.584	-0.20	(-0.69, 0.29)	-0.97	0.599	-0.17	(-0.48, 0.15)	-1.05	0.294
	Cohort 2 vs 3	0.49	(0.10, 0.87)	2.97	0.009	0.33	(-0.09, 0.77)	0.16	0.155	-	-	-	-
	Cohort 1 vs 3	0.34	(-0.03, 0.71)	2.19	0.074	0.13	(-0.35, 0.63)	0.79	0.789	-	-	-	-
		Anova test: F=4.57, p=0.011				Anova test: F=1.74, p=0.179				Anova test: F=1.11, p=0.294			

Appendix F

Paired t-test for each test comparison by cohort

		Test 1 vs Test 2					Test 2 vs Test 3					Test 1 vs Test 3				
Cohort	**Subcomponent**	**N**	**Mean diff.***	**(95% CI)**	**t statistic**	***p* value**	**N**	**Mean diff.**	**(95% CI)**	**t statistic**	***p* value**	**N**	**Mean diff.**	**(95% CI)**	**t statistic**	***p* value**
Cohort 1	**Vocabulary**	38	0.24	(0.05, 0.46)	2.48	0.018	37	0.08	(-0.08, 0.25)	1.00	0.324	77	0.38	(0.24, 0.51)	5.43	<0.001
	Structures	38	0.13	(-0.06, 0.32)	1.40	0.169	37	0.00	(-0.19, 0.19)	0.00	1.000	76	0.18	(0.02, 0.35)	2.22	0.03
	Reading	38	0.05	(-0.16, 0.27)	0.50	0.624	37	0.11	(-0.11, 0.33)	1.00	0.324	76	0.13	(-0.07, 0.33)	1.32	0.191
	Writing	38	0.05	(-0.20, 0.31)	0.42	0.676	37	0.27	(0.002, 0.54)	2.04	0.048	77	0.25	(0.09, 0.40)	3.13	0.003
	Average	38	0.12	(0.01, 0.23)	2.20	0.035	37	0.11	(-0.01, 0.24)	1.93	0.061	76	0.24	(0.14, 0.33)	4.89	<0.001
Cohort 2	**Vocabulary**	61	0.18	(0.05, 0.31)	2.82	0.007	53	0.08	(-0.11, 0.26)	0.81	0.419	61	0.21	(0.04, 0.38)	2.51	0.015
	Structures	61	0.02	(-0.21, 0.24)	0.14	0.885	53	-0.02	(-0.22, 0.19)	-0.20	0.855	61	0.05	(-0,13, 0.23)	0.55	0.582
	Reading	61	0.16	(-0.01, 0.34)	1.86	0.068	53	-0.11	(-0.32, 0.09)	-1.10	0.278	61	-0.02	(-0.24, 0.21)	-0.14	0.885
	Writing	61	-0.07	(-0.23, 0.10)	-0.78	0.437	53	0.17	(-0.02, 0.36)	1.77	0.083	61	0.10	(-0.10, 0.30)	0.97	0.335
	Average	61	0.07	(-0.02, 0.17)	1.52	0.135	53	0.03	(-0.09, 0.14)	0.49	0.624	61	0.09	(-0.01, 0.18)	1.89	0.064
Cohort 3	**Vocabulary**	57	-0.04	(-0.17, 0.10)	-0.53	0.597	-	-	-	-	-	-	-	-	-	-
	Structures	57	0.00	(-0.21, 0.21)	0.00	1.000	-	-	-	-	-	-	-	-	-	-
	Reading	57	-0.07	(-0.32, 0.18)	-0.56	0.576	-	-	-	-	-	-	-	-	-	-
	Writing	57	-0.26	(-0.51, -0.01)	-2.12	0.038	-	-	-	-	-	-	-	-	-	-
	Average	57	-0.09	(-0.24, 0.05)	-1.27	0.208	-	-	-	-	-	-	-	-	-	-

***Mean Difference**

Appendix G

Longitudinal Analysis of panel data

Subcomponent	N	Cohorts	Mean Improvement	95% CI	z-score	*p* value
Vocabulary	534	Cohort 1	0.181	(0.117, 0.245)	5.52	**<0.0001**
		Cohort 2	0.123	(0.051, 0.194)	3.37	**0.001**
		Cohort 3	-0.021	(-0.171, 0.130)	-0.27	0.788
Structures	532	Cohort 1	0.086	(0.004, 0.167)	2.05	**0.04**
		Cohort 2	0.025	(-0.065, 0.116)	0.55	0.583
		Cohort 3	0.005	(-0.186, 0.195)	0.05	0.961
Reading	533	Cohort 1	0.057	(-0.034, 0.148)	1.22	0.222
		Cohort 2	0.004	(-0.097, 0.105)	0.08	0.939
		Cohort 3	-0.056	(-0.268, 0.156)	-0.52	0.604
Writing	533	Cohort 1	0.113	(0.028, 0.199)	2.61	**0.009**
		Cohort 2	0.053	(-0.042, 0.148)	1.09	0.277
		Cohort 3	-0.258	(-0.458, -0.058)	-2.53	**0.011**
Average	2132	Cohort 1	0.116	(0.068, 0.163)	4.79	**<0.0001**
		Cohort 2	0.045	(-0.006, 0.098)	1.70	0.088
		Cohort 3	-0.087	(-0.197, 0.023)	-1.56	0.119

The table reports the results of a generalised least squares (GLS) regression with random effects, where tests 1 to 3 represent a time variable. The mean improvement between any test and the next as predicted by the model is shown along with 95% CIs and *p* values. Mean improvement was adjusted for both cohort and time-point, as well as any interactions between the two.

Appendix H

Possible trajectories of CEFR levels across time including number of students and percentage

Vocabulary	N	%	Structures	N	%	Reading	N	%	Writing	N	%
A1A1A1	1	1.11	A1A1A1	3	3.33	A1A1A1	16	17.78	A1A1A1	6	6.67
A1A2A2	1	1.11	A1A1A2	1	1.11	A1A1A2	4	4.44	A1A1A2	2	2.22
A1B1B1	1	1.11	A1A2A1	1	1.11	A1A2A1	2	2.22	A1A2A1	3	3.33
A2A2A2	4	4.44	A1B1B1	1	1.11	A1A2A2	3	3.33	A1A2A2	8	8.89
A2A2B1	7	7.78	A1B2B1	1	1.11	A1A2B1	1	1.11	A1A2B1	1	1.11
A2B1A2	4	4.44	A2A2A1	1	1.11	A2A1A1	2	2.22	A2A1A1	4	4.44
A2B1B1	4	4.44	A2A2A2	8	8.89	A2A1A2	5	5.56	A2A1A2	3	3.33
A2B1B2	1	1.11	A2A2B1	3	3.33	A2A2A1	4	4.44	A2A1B1	3	3.33
B1A2A2	1	1.11	A2B1B1	4	4.44	A2A2A2	9	10	A2A2A1	3	3.33
B1A2B1	1	1.11	B1A2A1	1	1.11	A2A2B1	5	5.56	A2A2A2	14	15.56
B1B1A2	2	2.22	B1A2A2	1	1.11	A2B1A1	1	1.11	A2A2B1	5	5.56
B1B1B1	16	17.8	B1A2B1	3	3.33	A2B1A2	3	3.33	A2B1A2	1	1.11
B1B1B2	6	6.67	B1B1A2	4	4.44	A2B1B1	4	4.44	A2B1B1	2	2.22
B1B2B1	4	4.44	B1B1B1	17	18.89	A2B1B2	1	1.11	B1A2A2	5	5.56
B1B2B2	4	4.44	B1B1B2	6	6.67	A2B2B2	1	1.11	B1A2B1	5	5.56
B2B1B1	2	2.22	B1B2B1	6	6.67	B1A2A1	1	1.11	B1A2B2	1	1.11
B2B1B2	1	1.11	B1B2B2	3	3.33	B1A2A2	2	2.22	B1B1A2	2	2.22
B2B2B2	23	25.6	B2A2B2	1	1.11	B1A2B1	1	1.11	B1B1B1	9	10
B2B2C1	1	1.11	B2B1B1	1	1.11	B1B1A2	1	1.11	B1B1B2	4	4.44
B2B2C2	1	1.11	B2B1B2	3	3.33	B1B1B1	9	10	B1B2B1	3	3.33
B2C1B2	1	1.11	B2B2B1	2	2.22	B1B1B2	1	1.11	B1B2B2	1	1.11
B2C1C1	2	2.22	B2B2B2	9	10	B1B2B1	2	2.22	B2B1B2	1	1.11
C1C1B2	1	1.11	B2B2C1	1	1.11	B2A2A2	1	1.11	B2B2C1	3	3.33
C1C1C1	1	1.11	B2C1B2	3	3.33	B2B1B2	1	1.11	B2C1B2	1	1.11
			C1B2B2	1	1.11	B2B2A2	1	1.11			
			C1C1B2	1	1.11	B2B2B1	1	1.11			
			C1C1C1	3	3.33	B2B2B2	2	2.22			
			C2B2B2	1	1.11	B2C1B1	1	1.11			
						B2C1C1	4	4.44			
						C1C2C1	1	1.11			
Total	90	100	Total	90	100	Total	90	100	Total	90	100

Appendix I: Sensitivity Analysis

Summary statistics for the 90 students that had valid scores across all three tests

		Test 1			Test 2			Test 3	
Subcomponent	**N**	**Mean**	**(95% CI)**	**N**	**Mean**	**(95% CI)**	**N**	**Mean**	**(95% CI)**
Vocabulary	90	3.1	(2.91, 3.29)	90	3.3	(3.12, 3.48)	90	3.38	(3.20, 3.56)
Structures	90	3.03	(2.82, 3.25)	90	3.12	(2.92, 3.33)	90	3.11	(2.91, 3.31)
Reading	90	2.18	(1.96, 2.39)	90	2.3	(2.05, 2.55)	90	2.28	(2.04, 2.51)
Writing	90	2.22	(2.04, 2.40)	90	2.2	(2.01, 2.39)	90	2.41	(2.20, 2.62)

Longitudinal regression results for the 90 students that had valid scores across all three tests

Subcomponent	**Mean Improvement**	**95% CI**	**z-score**	**p-value**
Vocabulary	0.14	(0.08, 0.20)	4.42	<0.001
Structures	0.04	(-0.03, 0.11)	1.05	0.295
Reading	0.05	(-0.03, 0.13)	1.29	0.196
Writing	0.09	(0.20, 0.17)	2.51	0.012