CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 7

Assessing writing proficiency in a Saudi Arabian university: Comparing students, tutors, and raters' assessment using selected CEFR scales

Ebtesam Abdulhaleem, King Salman Global Academy for Arabic Language, Riyadh, KSA

Claudia Harsch, Universität Bremen, Germany

https://doi.org/10.37546/JALTSIG.CEFR7-4

This study explores the writing proficiency levels of Saudi Arabian medical track students after completing a one-year Preparatory Year Programme (PYP), as well as the applicability of the Common European Framework of Reference for Languages (CEFR) in assessing their proficiency. The standardized writing exam administered at the end of the PYP revealed a ceiling effect, with the majority of students achieving high scores, despite the fact that the PYP teaches English at three different levels (beginner, intermediate, advanced). To obtain a more nuanced understanding of students' writing skills, alternative assessment methods were explored using selected CEFR scales, including self-assessment, tutor assessment, and assessment by raters recruited from the UK (experts in using CEFR scales). The study aimed to determine if these CEFR-based assessments can reliably differentiate among the three PYP levels, and if the CEFR scales are practical and applicable in this context. The findings show that the CEFR-based scores from all three assessor groups can reliably separate students according to their PYP level. The results highlight that the CEFR can serve as a valuable tool for understanding students' writing proficiency, even in non-European settings. This study encourages further exploration in the use of CEFR scales to assess proficiency levels.

Keywords: Writing proficiency, Preparatory Year Programme (PYP), Common European Framework of Reference for Languages (CEFR), self-assessment, tutors' assessment, raters, proficiency levels, CEFR scales

1 Introduction and background

Writing in English is a skill that many Saudi students find exceptionally challenging (McMullen 2009; Shukri 2014). This is true even among highly proficient students (Shukri 2014). To address this, Saudi Arabia has implemented Preparatory Year Programmes (PYPs) aimed at enhancing students' English skills during their initial year at university. These programmes aim to equip students with the necessary proficiency to navigate the English-medium academic environments of various colleges they will join after completing the PYP (Ebad 2014).

At the beginning of the PYP, students are grouped into three proficiency levels (elementary, intermediate or advanced) based on their test scores on the Oxford Placement Test (OPT) (OUP 2001), which evaluates students' listening and reading skills, along with grammar and vocabulary knowledge. However, the OPT does not assess written or oral skills, so proficiency in those areas remains unidentified prior to the PYP.

The OPT is scored between 0 and 100. Students scoring 0–45 are placed in the elementary level, those scoring 46–85 in the intermediate level, and those scoring above 85 in the advanced level.

At the end of the PYP, all students, regardless of the level they attend, take the same standardized proficiency exam, which includes a writing component. The exam only requires students to write a minimum of 120 words in 60 minutes on an easy, general descriptive topic about their daily routine at the university (see Appendix 1 for two performance examples, the exam itself cannot be published). It was designed based on a very low benchmark (roughly equivalent to CEFR level A2). The results of the exam revealed a ceiling effect, with scores concentrating at the upper end of the grading scale: 73% of all students achieved the highest score (10/10), regardless of the PYP level they had attended. The median and interquartile range (IQR) scores were 9.6 (9.2, 10), 10 (9.6, 10), and 10 (10, 10) for students starting the PYP at the elementary, intermediate, and advanced levels, respectively. While these high scores might indicate progress due to instruction during the PYP, or suggest that the exam was not adequately challenging, or had an insufficiently discriminating marking scheme, they do not effectively differentiate between students or accurately describe their proficiency according to an internationally recognized framework such as the CEFR. Consequently, determining the students' 'true' proficiency levels by the end of the programme proved to be challenging.

Methods that could be used to differentiate between students' levels may include assessments by the students themselves, by their teachers, or by independent raters. All methods may have advantages and disadvantages.

Self-assessment may be unreliable, since low-proficiency students tend to overestimate their proficiency (Babaii et al. 2016; Blue 1988; Leach 2012; Ünaldı 2016;). This has been described as the "metacognitive deficits" of the "Dunning-Kruger effect", i.e., it takes a certain level of competency to be able to assess one's own proficiency (Kruger and Dunning 1999). Self-assessment may also be inaccurate due to students' lack of experience in this approach (Babaii et al. 2016; Engelhardt and Pfingsthorn 2013).

Conversely, higher proficiency students may underestimate their own proficiency level (Kruger and Dunning 1999; Hodges et al. 2001; Lejk and Wyvill 2001; Tejeiro et al. 2012), possibly due to students being over-modest (Kun 2016). At the highest proficiency, researchers described more similarities between the students' and their teachers' assessment and therefore considered self-assessment as more accurate at higher-proficient levels (Kun 2016; Ünaldı 2016; Sahragard and Mallahi 2014).

As noted by Paris and Winograd (1990), familiarisation with and instruction in this approach can improve the accuracy and reliability of self-assessment. One way to determine the accuracy of self-assessment is to compare it with other methods, such as tutors' judgments or other test scores (Abdulhaleem and Harsch 2018; Ashton 2014; Babaii et al. 2016; Boud 1991), although high correlations between self-assessment and other measures of performance are unlikely (Dunning et al. 2004). For example, Falchikov and Boud (1989), in their meta-analysis of studies comparing self-assessment with teachers' marks, reported an average correlation of r=0.39. Correlation between self-assessment and students' - 'actual performance' (e.g., scores in a test) was very low (r=0.21) (Falchikov and Boud 1989).

In a similar way, teacher assessment may show comparably low correlations with scores allocated by external raters or with scores from standardized tests. Fleckenstein et al. (2018) found a correlation of r=0.41 between tutor assessments and test scores, noting that teachers overestimated students' levels compared to their actual performance in an achievement test. This overestimation was similarly evident in Bérešová's (2011) study, where teachers tended to overestimate students' vocabulary, grammar and language use compared with actual test results.

The CEFR proficiency framework has been employed to assess students' proficiency levels within Europe and beyond (e.g., Atai and Shoja 2011; Dragemark Oscarson 2009; Ünaldı 2016). Moreover, the CEFR is already used at the PYP curriculum, mainly to articulate the programme's objectives and to choose textbooks for each of the PYP levels. The principal reasons for the use of the CEFR in our study were the fact that it is already used in the PYP, the CEFR's design, and its role as a common metalanguage.

The CEFR "can be presented and exploited in a number of different formats, in varying degrees of detail" (Council of Europe [CoE] 2001: 36). The descriptors correspond well with the communicative teaching paradigm (Green 2012). Descriptors can "specify learning objectives in terms of situation, activities, functions and notions" (Green 2012: 21); and each descriptor "is worded in positive terms, even for lower levels" (North 2014: 55). The CEFR is used to "foster mutual understanding" across different users (Tannenbaum and Wylie 2005: 41); as a reference tool for identifying learners' needs prior to designing the curriculum (Little 2007); and as "a point of departure" (North 2014) to start the reflection, analysis and discussion of potential university standards and admission criteria (Harsch 2018). There are 53 CEFR scales representing different language skills and these must "be used selectively" (North 2014: 11) to suit the context in which they are employed.

2 Aim of the study

Although several studies have been conducted on Saudi students' writing skills in general (Aljumah 2012; Alkubaidi 2014; Hellmann 2013; McMullen 2009; McMullen 2014; Obeid 2017; Oraif 2016), to our knowledge, none has investigated the writing proficiency of Saudi medical track (MT) students in relation to the CEFR. The main objective of the study was therefore to obtain a more nuanced understanding of students' writing proficiency than the current exam upon completion of the three levels of the PYP-MT allows. Moreover, by comparing CEFR-based assessment from the perspectives of students and their tutors, we set out to explore the applicability of the CEFR in the Saudi Arabian PYP context, where the CEFR is not commonly used and where participants have not yet been thoroughly familiarised with this framework. Hence, students and their teachers assessed the end-of-year performances (from the standardised exam) against a CEFR-based assessment grid that contained selected CEFR writing scales. To triangulate the findings from within the PYP context, the same student performances were also assessed by external raters familiar with the CEFR, using the Writing Grid from the manual for relating language examinations to the CEFR (CoE 2009). We aimed to explore new ways of assessments that could reliably differentiate students (thus avoiding the aforementioned ceiling effect), while simultaneously benchmarking the three PYP levels against an internationally recognised framework (i.e., the CEFR). Hence, it was important to understand the extent to which scores given by students, their tutors and independent raters were comparable and correlated with each other.

Research questions

The study addresses the following research questions:

RQ1: Can students' self-assessment, tutors' assessment, and raters' assessment (using selected CEFR scales) reliably differentiate students' writing proficiency among the three PYP levels?

RQ2: To what extent are the scores from the three assessor groups comparable, taking the three PYP levels into account?

3 Methods

3.1 Overall design

The study takes a cross-sectional quantitative design. Three assessor groups assessed the same students' writing proficiency: students, their teachers and external raters. Students and their teachers assessed students' general writing proficiency, using similar assessment grids based on selected CEFR scales. Raters assessed the students' performances elicited by the end-of-year exam, using the CEFR grid from the Manual. The resulting scores from these three groups were quantitatively analyzed. The extent to which each group of assessors was able to discriminate reliably between the three PYP levels (RQ1) was analyzed using ANOVA and comparisons of means between levels, with pairwise comparisons between each pair of levels (elementary vs. intermediate vs. advanced). The scores obtained from all three assessor groups (RQ2) were compared between pairs (students vs. tutors vs. raters) using ANOVA and independent t-tests.

3.2 Participants

The study targeted female students in the PYP-MT, as they are being prepared to enter various medical and healthcare-related colleges such as the Colleges of Medicine, Pharmacy, Dentistry, Nursing, and Applied Medical Studies. The entire female cohort of students in PYP-MT (N=640) in 2016 was invited to participate, resulting in a total of n=517 participants across the three PYP levels (elementary, intermediate, and advanced). Of the participants, 90% were Saudi and 10% were non-Saudi, aged 18–19 years.

Furthermore, all PYP tutors (N=24) teaching English to the students in the PYP-MT were also offered the opportunity to participate, with a total of n=19 tutors accepting the invitation. All participating tutors were only teaching one level (either elementary, intermediate or advanced) when the data were collected, to try to reduce any 'norm-orientation' (comparison of a student with students in other levels) during data collection, although some tutors had previous teaching experience in teaching the other levels. The study analysis included a total of n=517 students whose general proficiency was assessed by both themselves and their tutors.

To examine students' and tutors' scores in relation to external measures, seven raters from two language institutes in the UK, who were experienced with writing assessment in higher education, familiar with the CEFR framework and experienced with using CEFR-based rating scales for rating second language texts, were invited to participate and accepted. They assessed the end-of-year performances by a subsample of 105 of the 517 students who participated in this study.

3.3 Ethics

Ethical permission was granted by the University of Warwick regarding the application, instruments and data collection (as part of a PhD study). Official permission was also given from the Dean of the PYP and the PYP research committee to collect data on the women's campus and to analyze the students' final exam written texts. All participants were fully informed about the aims of the research and the consequences of their participation (Punch 2005), and that it was possible to withdraw from the study at any time during or after participation; they were also given the chance to ask any questions regarding the study. All participants received an information sheet about the study, including all relevant contact information and a consent form to be signed. Both were translated into Arabic to ensure full comprehension.

3.4 Instruments

Due to administrative constraints, we were unable to provide students with a newly-developed exam specifically designed to operationalise the CEFR levels. Hence, we resorted to combining three different assessment perspectives, i.e., self-assessment, programme tutor assessment, and assessment by seven external raters. Students and tutors employed similar CEFR grids that were selected to analyse whether the student could achieve the writing construct in question (from their knowledge of themselves or the students); raters used the Assessment grid from the Manual to rate the same students' performances from the final exam.

For the student and tutor assessment grids, we selected the following ten CEFR scales relevant for assessing writing: *Overall Written Production, Overall Written Interaction, Type of Texts, What Can They Write, Vocabulary Range & Control, Grammatical Accuracy, Orthographic Control*, *Processing Texts, Reports and Essays* and *Note Taking*. Their relevance (face validity) to this study's context was checked with two teachers on the PYP and a member of academic staff working in one of the university medical colleges.

Irrelevant scales (e.g., *Correspondence* and *Creative writing*) were excluded as they are not related to the study's context. After designing the assessment grid and before piloting, more feedback was sought from the same teachers and from colleagues from the applied linguistic field. Based on this, further scales were either eliminated or combined, e.g., *Vocabulary range* and *Vocabulary control* were combined to reduce the burden on participants (Faez et al. 2011) and therefore increase the likelihood of their engagement in the assessment. Equally, however, there was a need to ensure that relevant writing scales were covered to gather a more complete picture of the students' writing levels.

In the assessment grids, the CEFR levels A1 to C2 (including plus levels for A2, B1, and B2) were depicted as columns 1 to 9; the 10 CEFR-based categories were described in 10 separate lines, with the respective descriptors located at their correct levels (see Appendix 2). Where the CEFR scales did not contain a descriptor for the plus level, we left a blank. This basic grid was then slightly amended for the student and tutor version.

3.4.1 Students' grid

For the student grid, the "can-do" descriptors were reformulated in "I can do" statements. Using the CEFR scales based on what learners "can do" with language (CoE 2001) may improve the reliability of the findings, as using functional language (i.e., "can do" statements) has been found to increase the accuracy of self-assessment (Ross 1998). For each descriptor, students were asked to decide whether they are confident that they can perform what is depicted in the descriptor ("Yes I can"), or whether they are "not sure" that they could perform the depicted language activity. We chose the "not sure" option to allow for doubts regarding students' abilities (Alderson 2005). When students choose "Yes I can", this, in the researchers' view (by adopting a more 'conservative' approach), indicates that students are most probably able to perform the language activity depicted in that descriptor. We decided against providing a third option (e.g., "cannot do"), as this would make the analysis more complex and difficult to interpret (Ashton 2014). Figure 1 shows how the grid works.

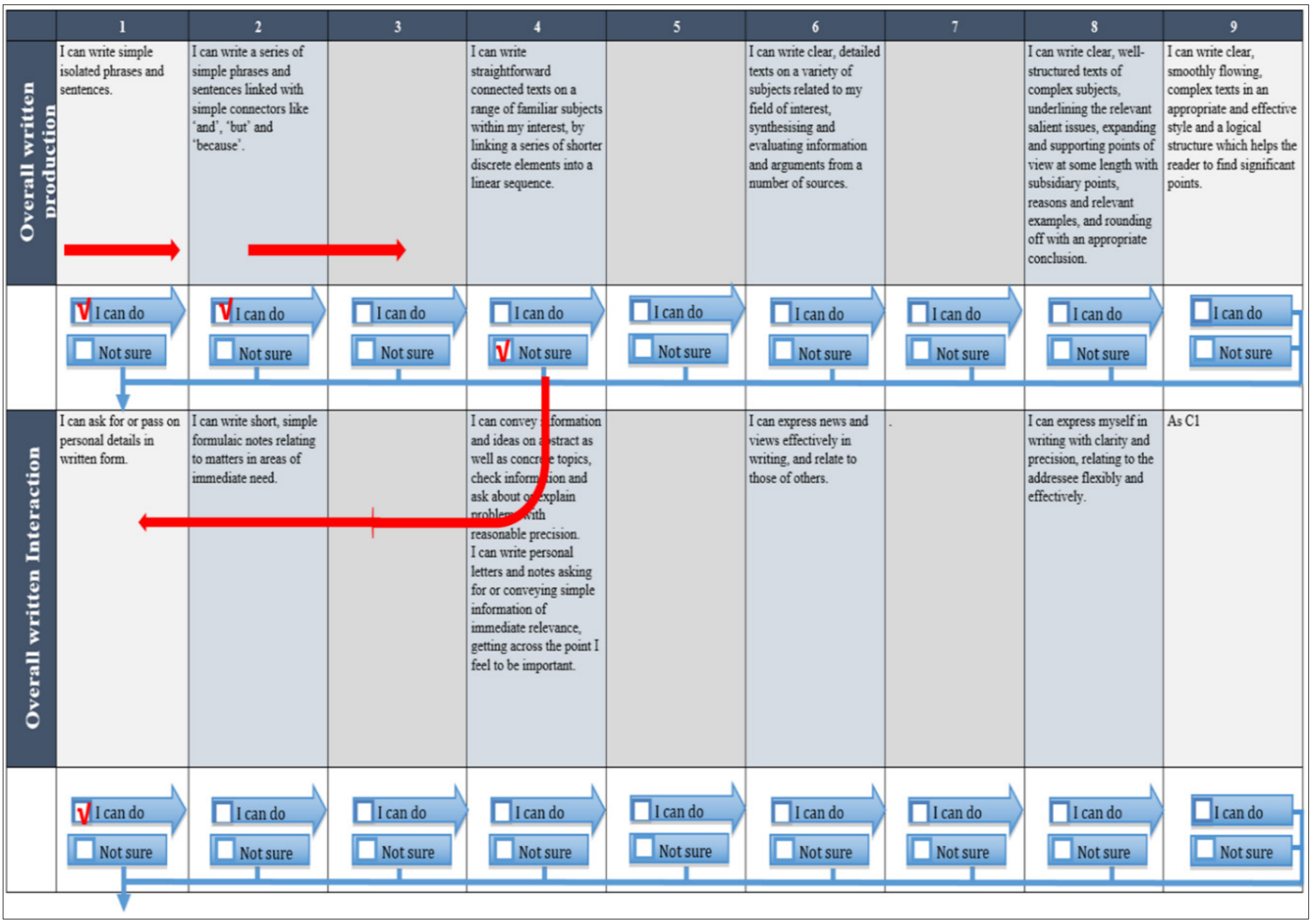
Students are required to read the descriptors starting with *Overall Written Production*, descriptor for level A1 (1 in the grid). If they feel they can do what the descriptor states, they tick "I can do" and move on to the second descriptor, and so on until they reach a descriptor that they feel they are not sure they are capable of doing or are unable to do. In the case in Figure 1, the student ticked not sure for the descriptor at level 4 (B1). In this case, the student then proceeds to the next row (i.e., the following CEFR-based category, here *Overall Written Interaction*) and follows the same process. The student's assessment for each category is coded as the last level at which they ticked "Yes I can", in the case above the student would score 2 (A2) for *Overall Written Production*, as there is no descriptor for level A2+.

3.4.2 Tutors' grid

The tutor grid was based on the same CEFR-based grid described above. The only difference to the student grid was, that the "can do" statements were rephrased as "The student can". Tutors used the same procedure as outlined above to assess each of their students.

3.4.3 Raters' grid

Raters used the *Writing Assessment Grid* from the CEFR manual mentioned earlier (CoE 2009: 187) to assess the aforementioned student performance. We did not adapt the Grid as we wanted to use it as an independent external criterion that should reflect the CEFR construct of writing as closely as possible. Hence, the raters used the grid in its original form, encompassing the six CEFR levels (A1 to C2, without plus levels) for the five categories *Overall, Range, Coherence, Accuracy,* and *Description*.



	1	2	3	4	5	6	7	8	9
Overall written production	I can write simple isolated phrases and sentences.	I can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.		I can write straightforward connected texts on a range of familiar subjects within my interest, by linking a series of shorter discrete elements into a linear sequence.		I can write clear, detailed texts on a variety of subjects related to my field of interest, synthesising and evaluating information and arguments from a number of sources.		I can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.	I can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.
	☑ I can do □ Not sure	☑ I can do □ Not sure	□ I can do □ Not sure	□ I can do ☑ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure
Overall written Interaction	I can ask for or pass on personal details in written form.	I can write short, simple formulaic notes relating to matters in areas of immediate need.		I can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision. I can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point I feel to be important.		I can express news and views effectively in writing, and relate to those of others.	.	I can express myself in writing with clarity and precision, relating to the addressee flexibly and effectively.	As C1
	☑ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure

Figure 1. Student assessment grid

3.5 Data collection

Data were collected during the final stages of the PYP year, after students had taken the final PYP exam, in the expectation that participants would have developed the necessary writing skills by then.

3.5.1 Students and their tutors

All students were given a study information sheet and were familiarized with the grid. The way the CEFR scales were formatted for this study aimed to help guide students in their self-assessment, and while there was no formal training conducted to improve the reliability of assessment (Harris 1997; Little 2002; Ross 1998) nor experience in self-assessment (Engelhardt and Pfingsthorn 2013), detailed instructions were given.

Each student received her own paper-based assessment grid bearing her name and university ID (Arabic version, anonymized after data collection), so that students could be tracked, and their assessments compared with those conducted by the tutors. To mitigate against the possibility of deliberately giving inaccurate assessments of their abilities, students were encouraged to assess themselves honestly; they were reassured that their assessment would not affect any of their marks and would only be used for research purposes.

Tutors were given the same study information sheet as the students and were familiarized with the grid before using it. They received one grid for each student, containing their names and university IDs.

A total of 517 students (73 elementary, 268 intermediate and 176 advanced) submitted self-assessments and were also rated by their teachers.

3.5.2 Raters

Raters received a two-hour training session that entailed familiarisation, standardisation and benchmarking activities adapted from the manual (CoE 2009) to use the CEFR grid. After training, each of the seven raters rated the same 105 texts (the aforementioned random sample of students' performances on the PYP end-of-year exam, the same performances that had been graded by the programme tutors which yielded the ceiling effect mentioned previously). Out of these 105 students, 14 attended the elementary level of the PYP, 55 the intermediate and 36 the advanced level. The raters used the assessment grid from the manual, which originally contains the six main CEFR levels; for the data collection here, to achieve comparability with the aforementioned 9-point grid, we asked the raters to also consider the plus levels, albeit without descriptors. Raters entered their chosen levels for the five categories in a prepared excel sheet that contained these nine levels and five categories.

3.6 Methods of Analysis

We compared the results of these three perspectives (self, tutor and rater's assessments) for reliability within and between the three groups of assessors and their capability to differentiate the three PYP levels.

Cronbach's alpha showed a high reliability (of α=0.88 and α=0.95 for students and tutors' assessment, respectively), showing that the scale items measured the same underlying construct and allowed the possibility of using average scores from the ten CEFR scales (Bland and Altman 1997).

Inter-rater reliability for the five categories of the rating scale for raters was measured using Cronbach's alpha, which was also >0.8, indicating good consistency between raters, allowing to average the seven scores for each category and student.

3.6.1 RQ1

Descriptive analyses were utilized to calculate the mean and standard deviation of students' self-assessments, tutors' assessments, and raters' scores for each CEFR-based category, to ascertain whether their respective ratings yielded differences in students' performances by PYP levels.

To examine whether the differences found in the descriptive analyses are indeed significant across the three PYP levels (elementary, intermediate, and advanced), we used a one-way analysis of variance (ANOVA), as ANOVA "looks for differences between groups which are not due to chance" (Green 2013: 107). Each group of assessors was separately examined. First, the homogeneity of variance was tested (Pallant 2013). In cases where the assumption of equal variances was violated, non-parametric analysis of variance tests (i.e., the so-called Brown-Forsythe and Welch Tests, see e.g. Green 2013) were conducted. A significance level (P-value) of less than 0.05 indicates a significant difference in mean scores across the three PYP levels. In addition, the ANOVA results report η^2, which is a measure of effect size (larger effect sizes reflecting larger differences; Miles and Shevlin 2001): values around 0.02 indicate "small", values around 0.13 "medium" and values above 0.26 "large" effect sizes (Cohen 1988).

To determine the significance between each pair of the three PYP levels, we conducted post-hoc tests. If the assumption of homogeneity was met, we performed Tukey's Honestly Significant Difference (HSD) test (Pallant 2013); otherwise, for heterogeneity of variances, we used Tamhane's T2 test (Green 2013).

3.6.2 RQ2

Self-assessments and tutors' assessments were compared using a paired sample t-test (Field 2009) to identify any significant differences between the different assessments of the same students; then, correlation and agreement analyses were conducted to examine the direction and the level of agreement between these two assessor groups. To observe the strength and direction of the relationship between students' and tutors' assessments, Spearman's correlation coefficient (r) was used. Values of r of 0.00-0.19 indicate "very weak" correlation; 0.20-0.39 "weak"; 0.40-0.59 "moderate"; 0.60-0.79 "strong" and 0.80-1.0 "very strong" correlation. Additionally, the weighted Cohen's Kappa coefficient (for ordinal data such as our scores; Cohen 1968) was used to measure the degree of exact agreement between students and tutors, which takes into account the agreement that can be attributed to chance (Smeeton 1985). Kappa values of 0–0.2 indicate "slight" agreement, 0.21–0.4 "fair", 0.41–0.6 "moderate", 0.61–0.8 "substantial", 0.81–1 "almost perfect" and 1 "perfect" agreement (Landis and Koch 1977). In addition, percentages of exact agreement of student-tutor pairs were calculated, as well as agreement within one and within two adjacent CEFR levels.

For the 105 cases where three sets of data existed, we performed ANOVA, correlation and post-hoc tests, to compare the means of the self-assessments, tutors' assessments, and scores given by the external raters for the same students. This allowed for the examination of the direction and relation among the assessments provided by these three groups.

4 Results

4.1 RQ1 CEFR writing levels assessed by students, tutors, and raters separately across the three PYP levels

4.1.1. Descriptive Analysis

First, we present the results of the descriptive analyses (mean and standard deviation (SD)) for the three PYP levels (elementary, intermediate, and advanced), as perceived by students' self-assessment, tutors' and raters' assessments. Table 1 illustrates the self-assessment results, the results for tutors and raters are presented in Appendix 3 for space reasons.

Table 1. ***Descriptive analysis of PYP students' self-assessment across the PYP levels***

	Elementary n=73		Intermediate n=268		Advanced n=176	
CEFR Categories	M	SD	M	SD	M	SD
Overall Written Production	5.57	2.35	6.24	2.17	7.91	1.66
Overall Written Interaction	3.93	2.10	4.22	2.28	6.67	2.56
Type of Texts	3.94	2.05	4.28	2.23	6.27	2.48
What Can They Write	4.40	2.24	4.87	2.25	6.80	1.97
Vocabulary Range & Control	3.55	2.00	3.95	1.97	5.85	2.37
Grammatical Accuracy	4.32	2.68	5.08	2.39	6.16	2.84
Orthographic Control	5.05	2.77	5.41	2.67	7.00	2.14
Processing Texts	3.81	1.54	4.39	1.76	6.13	2.23
Reports and Essays	4.14	2.44	4.50	2.41	6.75	2.04
Note Taking	5.22	2.48	5.44	2.30	6.94	2.17
Average of Scales	4.48	1.58	4.92	1.53	6.73	1.43
M=Mean, SD=Standard deviation						
Coding scheme for CEFR categories: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)						

For each category and each group of assessors, mean scores increased from elementary to intermediate to advanced level students, indicating that the three group of assessors could differentiate between the three PYP levels, unlike the end-of-year exam.

4.1.2. ANOVA

To find out whether the increase across the three PYP levels is significant, we conducted ANOVA analyses. While we present the results for the three assessor groups here, the supporting tables are presented in the appendix for space reasons: Appendix 4 contains the tables for students; Appendix 5 for tutors and Appendix 6 for raters.

Looking at the students' self-assessment across the three PYP levels (Appendix 4, Tables 6 [ANOVA] and 7 [non-parametric analysis of variance tests]), the effect sizes were 0.095 to 0.26, indicating medium-to-large effect sizes for the differences between elementary, intermediate and advanced groups. The largest effect size was observed for the average of all categories (η^2=0.26). From the post hoc pairwise results (Appendix 4, Tables 8 [Tukey] and 9 [Tamhane]), significant differences were evident between the advanced and intermediate levels and the advanced and elementary levels. There were no significant differences between the elementary and intermediate levels, except in the *Processing Texts* category, where the scores for students from all three levels differed significantly from each other.

With regard to tutors' assessment, there were significant differences for all CEFR categories across the three PYP levels (Appendix 5, Tables 10 and 11). A substantial effect (η^2) was observed in most categories, except for *Note Taking*, where the effect was comparatively small. The results of the post-hoc tests (Appendix 5, Tables 12 and 13) showed significant differences in tutors' assessments between all three PYP levels, in the expected directions, with the elementary level receiving significantly lower scores compared to the intermediate level, and the intermediate level significantly lower than the advanced level.

When it comes to the external raters, we used the average scores across the seven raters (Appendix 6). The ANOVA (Table 14) showed significant differences across the three PYP levels, with large effect sizes in the expected directions (i.e., the elementary level receiving significantly lower scores compared to the intermediate level, and the intermediate level scoring significantly lower than the advanced level). The post-hoc analysis (Appendix 6, Table 15) showed significant differences in the raters' scores of students at the advanced versus intermediate or elementary levels for all categories (*Range, Coherence, Accuracy, Description* and *Overall*), but not between the intermediate and elementary levels in any category.

4.2 RQ2 comparing the three participating assessor groups: students, tutors and raters

RQ2 examined the extent to which the three participating assessor groups (students, tutors, raters) are comparable in their assessment using the selected CEFR-based categories. As two groups (students and tutors) used the same tool for assessment, we first compared these two groups, using a paired sample t-test to check whether the PYP students' and tutors' assessments differed significantly. Then, a comparison across the three groups was conducted, using correlations and ANOVA to compare the means between self-, tutors' and raters' scores of the same 105 students.

4.2.1. Self- and tutors' assessments

We compared means for students and tutors using the paired t-test. Cohen's d provides an estimate of the effect size (Pallant 2013), where d=0.2 is considered "small", 0.5 "medium" and 0.8 "large" (Cohen 1988). Appendix 7, Table 16 contains the detailed results.

At the elementary level, the largest effect sizes were observed for *Overall written production* and *Processing texts*, followed by *Note taking*, with students rating themselves significantly higher than their tutors. At the intermediate level, the largest (medium size) differences were for *Type of texts, Overall written interaction*, and *Vocabulary range and control*; in each case the students rated themselves lower than the tutors.

With the advanced-level students, scores on most of the CEFR-based categories showed very similar means (with non-significant P-values and small effect sizes), indicating that students and their tutors have similar perceptions of the CEFR levels students have reached in those categories. However, this was not true for all scales, with tutors scoring significantly higher for *Type of texts* and significantly lower for *Note taking* and *Reports and essays* (small effect size).

Appendix 7, Table 17 shows the correlation between students and teachers' scores, the weighted kappa (measure of agreement) and the percentages of scores with exact agreement (identical level assigned), or agreements within one or two levels.

There was a significant positive correlation between the scores of students and their tutors for all CEFR-based categories, though the strength of the relation was weak to moderate (all r<0.30 for individual items; r=0.39 for overall average). Weighted Kappa was low (max=0.39), indicating only weak to moderate agreement in students' and tutors' assessment. Overall, 19.0% of pairs agreed exactly; 52.4% agreed within one level and 79.9% within two levels, showing fairly close agreement between the tutors' assessment and their students' self-assessment.

4.2.2. Self-, tutors' and raters' assessments

Students', tutors' and raters' assessments were compared using only the sample where data exist from self-assessment, tutor assessment and mean scores across the seven raters (n=105, including all three levels). Correlation analysis was carried out to explore the relations between the three assessments (students, tutors and raters). Table 2 presents the results.

Table 2. ***Overall correlation analysis between self, tutors' and raters' assessment***

	Raters n=105		Students n=105		Tutors n=105	
	Pearson Correlation	P-value	Pearson Correlation	P-value	Pearson Correlation	P-value
Raters	1		0.44**	<0.001	-0.11	0.27
Students	0.44**	<0.001	1		-0.065	0.51
Tutors	-0.11	0.27	-0.065	0.51	1	

** Correlation is significant at the .01 level (2-tailed).

There exists a significant positive correlation between the raters' scores and the students' self-assessment, although the tutor's scores did not correlate significantly with either students' or raters' scores. The patterns of averaged scores from self-assessment for elementary, intermediate and advanced levels were B1, B1 and B2 (i.e., elementary and intermediate scored the same, then up one level for advanced students), and for raters the pattern was similar: A2+, A2+ and B1 (i.e., elementary and intermediate scored the same, then up one level for advanced students). However, the pattern for teachers' ratings differed: A2+, B1 and B2, respectively.

To compare the three groups, a one-way ANOVA was used. Table 3 shows the ANOVA results, comparing students', tutors' and raters' assessments for the 105 participants for whom all three types of assessments were available.

Table 3. *One-way ANOVA between students, tutors and raters*

	SS	df	MS	F	P-value	η^2
Between Groups	113.74	2	56.87	25.99	p<.0001	0.20
Within Groups	662.9	303	2.19			
Total	776.65	305				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η2=Effect size:0.02=small; 0.13=medium; 0.26=large.

We found significant differences in the scores between self-assessment, tutors' and raters' assessments, with a large effect size. To identify where the differences were located, Tukey's post hoc analysis was conducted (Appendix 8, Table 18), which showed that the raters gave significantly lower scores than both the students and tutors, and this was true across all three PYP levels (Appendix 8, Table 19).

5 Discussion and conclusion

This study aimed to explore assessments by three groups of assessors, i.e., students, their tutors and external raters, in order to yield assessment approaches that would be able to differentiate between the three proficiency levels taught at the PYP (Intensive English) programme for medical students. At the same time, we sought to benchmark the three PYP proficiency levels achieved in writing at the end of the PYP to a recognized framework (the CEFR). We also aimed to deepen our understanding of self-assessment, tutor assessment, and scores of independent raters based on relevant CEFR scales in the Saudi Arabian higher education context.

5.1 Research Question 1

Our first research question, i.e., "Can students' self-assessment, tutors' assessment, and raters' assessment (using selected CEFR scales) reliably differentiate students' writing proficiency among the three PYP levels?" was partially supported. The students placed in elementary level generally received lower scores compared to those at the intermediate level, and the intermediate level students scored lower than the advanced level students; differences were significant between advanced and intermediate students, and between advanced and elementary students, although the differences between elementary and intermediate students were less pronounced.

The CEFR can potentially be used to gain a criterion-referenced general overview of the students' proficiency levels as a starting point in a context outside of Europe such as Saudi Arabia, with participants having no or little experience with using the CEFR scales (Abdulhaleem and Harsch 2018). The scores could be benchmarked against a recognised framework (i.e., the CEFR), although only selected scales of the CEFR were used in the assessment grids. Scores for elementary, intermediate and advanced level students' self-assessments were equivalent to CEFR levels B1, B1 and B2; scores from tutor assessment placed students at A2+, B1+ and B2 respectively, while the external raters placed students at A2+, A2+ and B1. We will discuss the meaning of these results below, when taking a closer look at agreement levels.

5.2 Research Question 2

Our second research question, i.e., "To what extent are the scores from the three assessor groups comparable, taking the three PYP levels into account?" was also partially supported.

When comparing **students and tutors**, a moderate yet significant correlation between the students' self-assessments and tutors' assessments was found (r=0.39). This is similar to the average correlation identified by Falchikov and Boud (1989), in their meta-analysis of studies comparing self-assessment with teachers' marks, which also reported an average correlation of r=0.39.

Even if results correlate significantly, this does not necessarily demonstrate exact or close agreement (Fleiss and Cohen 1973; Cohen 1968). To the best of our knowledge, few studies investigating self-assessment – especially language proficiency-related studies – have compared agreement between students' self-assessment and their tutors' assessment. In this study, we used a weighted kappa to test the significance and percentage agreement between the two assessments. Exact agreement between students' and tutors' assessment was low (19%) but was higher between one (52.4%) and two (79.9%) adjacent CEFR scores. The two adjacent scores in the study means that the agreement is equal to "one and a half levels, e.g., A2+ to B1+", which is considered sufficient agreement according to the CEFR manual (CoE 2009: 37). This means that the students were not too far away in their perceptions of their CEFR levels from those of their tutors, suggesting the value of using the CEFR scales as exemplified in this study.

Looking at the three PYP proficiency levels separately, elementary students self-assessed their CEFR levels as B1, tutors assessed them as A2+. So elementary-level students tend to overestimate their proficiency (CEFR levels) compared to tutors. This was expected, as it has been widely found in the literature that low-proficiency students tend to overestimate their proficiency (Babaii et al. 2016; Leach 2012; Ünaldı 2016; Blue 1988).

Intermediate students achieved levels of B1 by self-assessment and B1+ by tutors. In contrast to the elementary level students, some intermediate-level students were found to underestimate their proficiency compared to their tutors' assessment. Similar results were also found in the literature, where higher proficiency students show a tendency to underestimate their proficiency level when they assess themselves (Kruger and Dunning 1999; Hodges et al. 2001; Lejk and Wyvill 2001; Tejeiro et al. 2012).

Advanced-level students achieved B2 according to self- and tutor-assessment. Generally, their self-assessment was closer to that of their tutors and showed less variance than at the other levels, indicating more accurate self-assessment. This was found in other studies that described more similarities between the students and their teachers' marks/assessment and therefore considered the assessment as more accurate when students came from higher-proficient levels (Kun 2016; Sahragard and Mallahi 2014; Ünaldı 2016), possibly due to the Dunning-Kruger effect, where students at higher proficiency levels have the cognitive ability to assess and judge their proficiency more accurately**.**

With regard to comparing **students and raters**, there was a significant moderate correlation between the students' self-assessments' and raters' assessments (r=0.44). The pattern of levels assigned by students at each of the proficiency levels (B1 and B2 for elementary, intermediate and advanced) was similar to that assigned by the raters (A2+, A2+ and B1, respectively), although the raters' assessments were around one CEFR level lower than the students' assessments across all PYP proficiency levels. These findings are consistent with those of Fleckenstein et al. (2018).

Comparing **tutors and raters**, agreement between these two groups was lower than between students and teachers or students and raters. Different explanations can be given for the discrepancies between the tutors' assessment and the raters' scores. One explanation is that though the tutors are following criterion-referenced assessment as it is usually the case when using the CEFR scales (Fleckenstein et al. 2018; Hughes 2002), there is still the possibility that the tutors tended to compare the students within or between their classes (norm-referenced assessment) (Fleckenstein et al. 2018; Lok et al. 2016). However, the grades assigned by the tutors were the most discriminating (different average CEFR levels assigned to elementary, intermediate and advanced level students), whereas students and raters gave the same levels to elementary and intermediate students.

Moreover, the raters were focusing on a small sample of specific exam texts, which may be easier to judge than students' proficiency in general (as for students and tutors using the CEFR scales) (Fleckenstein et al. 2018; Südkamp et al. 2012), However, raters only scored the end-of-year exam texts, which could have been inadequate to demonstrate students' full range of writing proficiency, as for example, level C1 requires complex subjects, a wide range of topics and imaginative texts, whereas the exam (based at A2 level) only required students to write 120 words in 60 minutes on a general topic about their daily routine at the university, with little scope to demonstrate higher skills. There may be a difference between what students and their teachers assess they “can do” in general and what they actually were able to demonstrate in the exam. Another source of variance is to be found in the grid the raters used, which may have been inappropriate for the exam at hand, or the rater training may have been inadequate.

5.3 Conclusions

Based on our findings, and despite the limitations identified, there are indications enough to argue for the usefulness of the CEFR to identify students’ proficiency levels. Students and tutors could potentially use the CEFR-based grids and compare their respective assessments as a basis for identifying areas on which to focus for further learning. Considering that the participating students and tutors had not been extensively trained in using the CEFR scales to identify students’ proficiency levels in writing, the findings for correlations and underestimation and overestimation of self-assessment are similar to those found in the literature. As mentioned in Moonen et al. (2013), many people have little experience of and exposure to the use of the CEFR scales, and as suggested by Davis (2015), Fahim and Bijani (2011), Fleckenstein et al. (2018), and Weigle (1994), with proper instruction and training, the tutors and students might be more accurate in their assessment.

The study findings revealed noticeable variations in the average scores across the three PYP levels in the assessments conducted by students, tutors, and raters. These disparities provide insights into the applicability of the CEFR scales. Furthermore, the results highlight that the CEFR can serve as a valuable criterion-referenced tool for gaining a broad understanding of students’ writing proficiency levels, even within a non-European setting where participants may possess limited familiarity with the CEFR scales. This serves as a foundation for future assessment and evaluation endeavors, encouraging further exploration in this area.

6 References

Abdulhaleem, Ebtesam & Claudia Harsch. 2018. Using the CEFR scales to assess students’ proficiency levels in a Saudi-Arabian higher education context. In Anikó Brandt, Astrid Buschmann-Göbels & Claudia Harsch. (eds.). *Der Gemeinsame Europäische Referenzrahmen für Sprachen und seine Adaption im Hochschulkontext. Erträge des 6. Bremer Symposions*. Fremdsprachen in Lehre und Forschung Bd. 51. [*The Common European Framework of Reference for Languages and its adaptation for higher education contexts. Proceedings of the 6th Bremer Symposium*. Foreign Languages in Teaching and Research 51.] Bochum: AKS. 167-178.

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

Aljumah, Fahad Hamad. 2012. Saudi learner perceptions and attitudes towards the use of blogs in teaching English writing course for EFL majors at Qassim University. *English Language Teaching* 5(1). 100-116. https://doi.org/10.5539/elt.v5n1p100.

Alkubaidi, Miriam A. 2014. The relationship between Saudi English major university students’ writing performance and their learning style and strategy use. *English Language Teaching* 7(4). 83-95. https://doi.org/10.5539/elt.v7n4p83.

Ashton, Karen. 2014. Using self-assessment to compare learners’ reading proficiency in a multilingual assessment framework. *System*, 42, 105–119. https://doi.org/10.1016/j.system.2013.11.006.

Atai, Mahmood Reza & Leila Shoja. 2011. A triangulated study of academic language needs of Iranian students of computer engineering: Are the courses on track? *RELC Journal* 42(3). 305-323. https://doi.org/10.1177/0033688211419392.

Babaii, Esmat, Shahin Taghaddomi & Roya Pashmforoosh. 2016. Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing* 33(3). 411-437. https://doi.org/10.1177/0265532215590847.

Bérešová, Jana. 2011. The impact of the common European framework of reference on teaching and testing in Central and Eastern European context. *Synergies Europe* 6. 177-190. https://gerflint.fr/Base/Europe6/jana.pdf (accessed 22 January 2025).

Bland, J. Martin & Douglas G. Altman. 1997. Statistics notes: Cronbach's alpha. *The British Medical Journal*, 314, 572. https://doi.org/10.1136/bmj.314.7080.572.

Blue, George M. 1988. Self-assessment: The limits of learner independence. Individualization and autonomy in language learning. *ELT Documents* 131. 100-118.

Boud, David. 1991. *Implementing student self-assessment.* Campbelltown, N. S. W.: Higher Education Research and Development Society of Australasia (HERDSA).

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4). 213–220. https://doi.org/10.1037/h0026256. PMID 19673146

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe. 2009. *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR).* Strasbourg: Council of Europe.

Davis, Larry. 2015. The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33(1). 117-135. https://doi.org/10.1177/0265532215582282.

Dragemark-Oscarson, Anne 2009. *Self-assessment of writing in learning English as a foreign language. A study at the upper secondary school level.* (Doctoral thesis, University of Gothenburg). http://hdl.handle.net/2077/19783 (accessed 22 January 2025).

Dunning, David, Chip Heath & Jerry M. Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest* 5(3). 69-106. https://doi.org/10.1111/j.1529-1006.2004.00018.x.

Ebad, Ryhan. 2014. The role and impact of English as a language and a medium of instruction in Saudi higher education institutions: Students-instructors perspective. *Studies in English Language Teaching* 2(2). 140-148. https://doi.org/10.22158/selt.v2n2p140.

Engelhardt, Maike & Joanna Pfingsthorn. 2013. Self-assessment and placement tests – a worthwhile combination? *Language Learning in Higher Education* 2(1). 75-89. https://doi.org/10.1515/cercles-2012-0005.

Faez, Farahnaz, Shelley Taylor, Suzanne Majhanovich, Patrick Brown & Maureen Smith. 2011. Teachers' reactions to CEFR's task-based approach for FSL classrooms. *Synergies Europe* 6. 109-120. https://gerflint.fr/Base/Europe6/faez.pdf (accessed 22 January 2025).

Fahim, Mansoor & Houman Bijani. 2011. The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing* 1(1). 1-16.

Falchikov, Nancy & David Boud. 1989. Student self-assessment in higher education: A meta-analysis. *Review of Educational Research* 59. 395–430. https://doi.org/10.3102/00346543059004395.

Field, Andy. 2009. *Discovering statistics using SPSS* (3rd ed.). Los Angeles: Sage Publications.

Fleckenstein, Johanna, Michael Leucht & Olaf Köller. 2018. Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly* 15(1). 90-101. https://doi.org/10.1080/15434303.2017.1421956.

Fleiss, Joseph L. & Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3). 613-619. https://doi.org/10.1177/001316447303300309.

Green, Anthony. 2012. *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range* (Vol. 2). Cambridge: Cambridge University Press.

Green, Rita. 2013. *Statistical analyses for language testers.* Basingstoke, UK: Palgrave Macmillan.

Harris, Michael. 1997. Self-assessment of language learning in formal settings. *ELT Journal* 51(1). 12–20. https://doi.org/10.1093/elt/51.1.12.

Harsch, Claudia. 2018. How suitable is the CEFR for setting university entrance standards? *Language Assessment Quarterly* 15(1). 102-108. https://doi.org/10.1080/15434303.2017.1420793.

Hellmann, Kate. 2013. *What do I need to succeed: The case of Arab engineering graduate students' self-perceptions of writing* (Doctoral dissertation, University of Idaho).

Hodges, Brian, Glenn Regehr & Dawn Martin. 2001. Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine* 76(10). S87-S89. https://doi.org/10.1097/00001888-200110001-00029.

Hughes, Arthur. 2002. *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Kruger, Justin & David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6). 1121-1134. https://doi.org/10.1037//0022-3514.77.6.1121.

Kun, András István. 2016. A comparison of self versus tutor assessment among Hungarian undergraduate business students. *Assessment & Evaluation in Higher Education* 41(3). 350-367. https://doi.org/10.1080/02602938.2015.1011602.

Leach, Linda. 2012. Optional self-assessment: Some tensions and dilemmas. *Assessment & Evaluation in Higher Education* 37(2). 137-147. https://doi.org/10.1080/02602938.2010.515013.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159-174. https://doi.org/10.2307/2529310.

Lejk, Mark & Michael Wyvill. 2001. The effect of the inclusion of self-assessment with peer assessment of contributions to a group project: A quantitative study of secret and agreed assessments. *Assessment & Evaluation in Higher Education* 26(6). 551-561. https://doi.org/10.1080/02602930120093887.

Levene, H. 1960. Robust tests for equality of variance. In Ingram Olkin (ed.), *Contributions to Probability and Statistics,* 278–292. Palo Alto, CA: Stanford University Press.

Little, David. 2002. The European language portfolio: Structure, origins, implementation and challenges. *Language Teaching* 35(3). 182-189. https://doi.org/10.1017/S0261444802001805.

Little, David. 2007. The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *Modern Language Journal* 91(4). 645-655. https://doi.org/10.1111/j.1540-4781.2007.00627.x.

Lok, Beatrice, Carmel McNaught & Kenneth Young. 2016. Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education* 41(3). 450-465. https://doi.org/10.1080/02602938.2015.1022136.

McMullen, Maram George. 2009. Using language learning strategies to improve the writing skills of Saudi EFL students: Will it really work? *System: An International Journal of Educational Technology and Applied Linguistics* 37(3). 418-433. https://doi.org/10.1016/j.system.2009.05.001.

McMullen, Maram George. 2014. The value and attributes of an effective preparatory English program: Perceptions of Saudi university students. *English Language Teaching* 7(7). 131-140. https://doi.org/10.5539/elt.v7n7p131.

Miles, Jeremy & Mark Shevlin. 2001. *Applying regression and correlation: A guide for students and researchers.* London: Sage Publications Ltd.

Moonen, Machteld, Evelien Stoutjesdijk, Rick de Graaff & Alexxandra Corda. 2013. Implementing the CEFR in secondary education: Impact on FL teachers' educational and assessment practice. *International Journal of Applied Linguistics* 23(2). 226–246. https://doi.org/10.1111/ijal.12000.

North, Brian. 2014. *The CEFR in practice.* Cambridge: Cambridge University Press.

Obeid, Rana. 2017. Second language writing and assessment: Voices from within the Saudi EFL context. *English Language Teaching* 10(6). 174-181. https://doi.org/10.5539/elt.v10n6p174.

Oraif, Iman. 2016. The right approach in practice: A discussion of the applicability of EFL writing practices in a Saudi context. *English Language Teaching* 9(7). 97-102. https://doi.org/10.5539/elt.v9n7p97.

Oxford University Press. 2001. *Quick placement test: Paper and pen test, user manual*.

Pallant, Julie. 2013. *SPSS survival manual* (5th ed.). Maidenhead, UK: McGraw-Hill Education.

Paris, Scott G. & Peter Winograd. 1990. How metacognition can promote academic learning and instruction. In Beau Fly Jones & Lorna Idol. (eds.). *Dimensions of Thinking and Cognitive Instruction,* 15-51. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Punch, Keith F. 2005. *Introduction to Social Research: Quantitative and Qualitative Approaches* (2nd ed.). London: Sage Publications Ltd.

Ross, Steven. 1998. Self-assessment in Second Language Testing: A Meta-analysis and Analysis of Experiential Factors. *Language Testing* 15(1). 1-20. https://doi.org/10.1177/026553229801500101.

Sahragard, Rahman & Mallahi, Omid. 2014. Relationship between Iranian EFL learners' language learning styles, writing proficiency and self-assessment. *Procedia-Social and Behavioral Sciences* 98. 1611-1620.

Shukri, Nadia Ahmad. 2014. Second language writing and culture: Issues and challenges from the Saudi learners' perspective. Arab World English Journal, 5(3), 190–207. https://awej.org/second-language-writing-and-culture-issues-and-challenges-from-the-saudi-arners-perspective (accessed 22 January 2025).

Smeeton, Nigel C. 1985. Early history of the kappa statistic. *Biometrics* 41(3). 795.

Südkamp, Anna, Johanna Kaiser & Jens Möller. 2012. Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology* 104(3). 743-762. https://doi.org/10.1037/a0027627.

Tannenbaum, Richard J., & E. Caroline Wylie. 2005. Mapping English language proficiency test scores onto the common European framework. *ETS Research Report Series* RR-05-18, TOEFL-RR-80. Princeton, NJ: Educational Testing Service.

Tejeiro, Ricardo A., Jorge L. Gómez-Vallecillo, Antonio F. Romero, Manuel Pelegrina, Agustín Wallace & Enrique Emberley. 2012. Summative self-assessment in higher education: Implications of its counting towards the final mark. *Electronic Journal of Research in Educational Psychology* 10(2). 789-812.

Ünaldı, İhsan. 2016. Self and teacher assessment as predictors of proficiency levels of Turkish EFL learners. *Assessment & Evaluation in Higher Education* 41(1). 67-80. https://doi.org/10.1080/02602938.2014.980223.

Weigle, Sara Cushing. 1994. Effects of training on raters of ESL compositions. *Language Testing* 11(2). 197-223. https://doi.org/10.1177/026553229401100206.

7 Biography

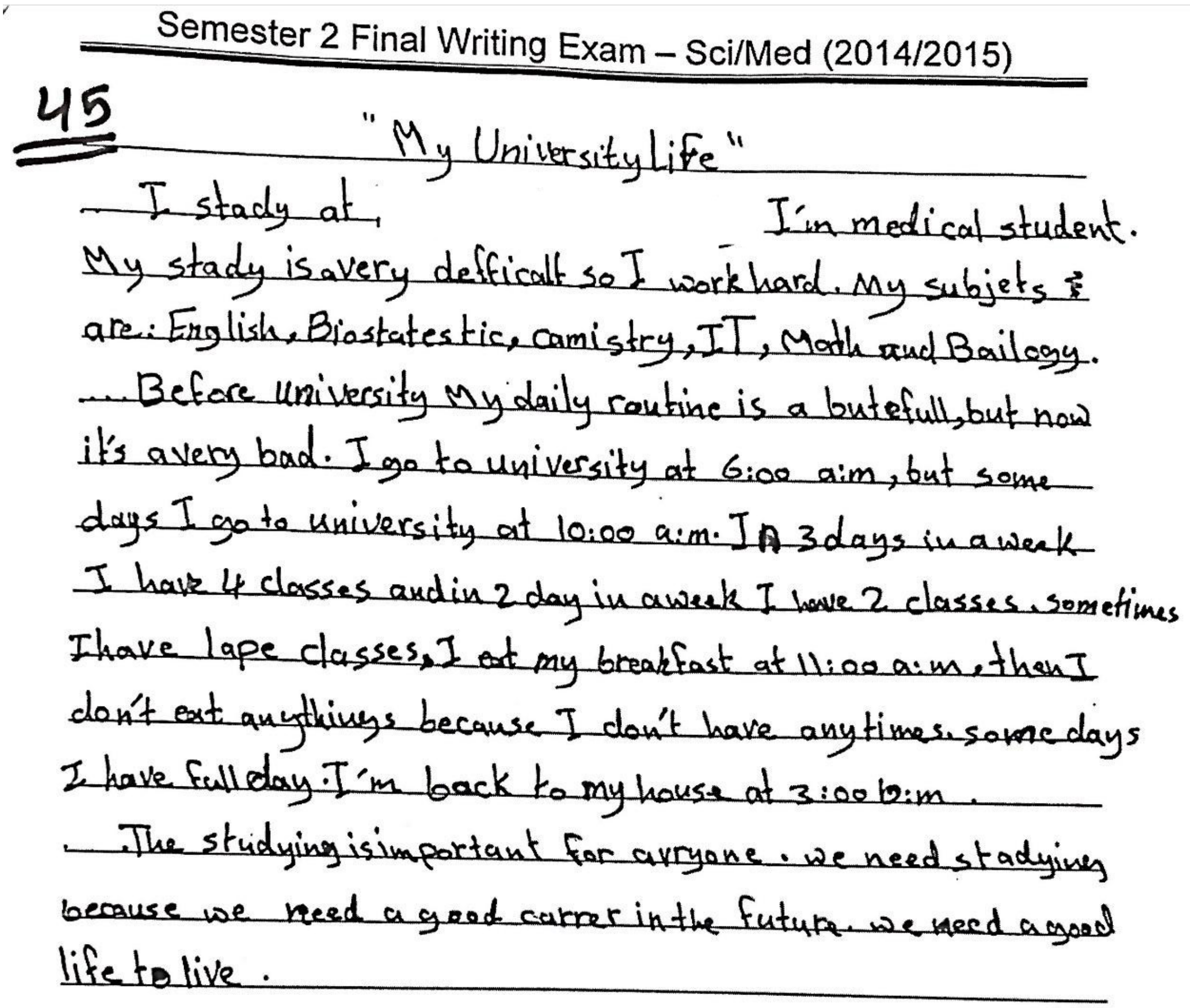
Ebtesam Abdulhaleem is an Assistant Professor and a former staff member at King Saud University. Currently, she is the Head of the Testing Development Unit at King Salman Global Academy for Arabic Language. Abdulhaleem holds a PhD in Applied Linguistics from the University of Warwick and a postgraduate certificate in language testing from Lancaster University. Her research interests encompass various areas, including language testing and assessment, assessment literacy, corpus linguistics, the role of practitioners as researchers, as well as well-being and professional development.

Claudia Harsch is a professor at the University of Bremen, specialising in language learning, teaching and assessment. She has worked in Germany and in the UK, and is active in teacher training worldwide.

Her research interests focus on areas such as language assessment, language and migration, the development of language assessment literacy, and the implementation of the CEFR. Claudia is currently the immediate past president of the International Language Testing Association (president from 2023-24), and was president of the European Association of Language Testing and Assessment from 2016-2019.

8 Appendices

8.1 Appendix 1: Samples of students' written texts in the end of year exam



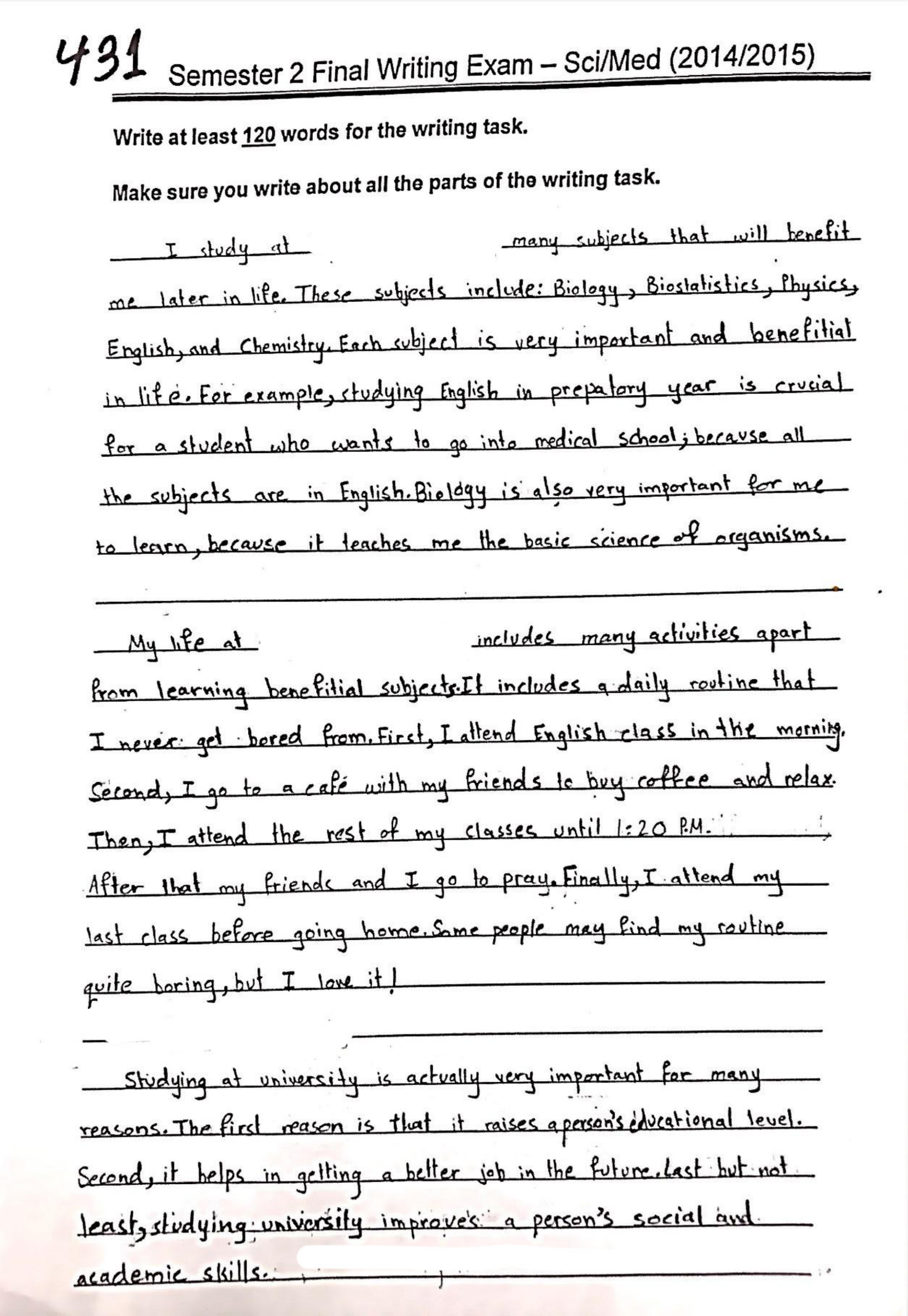
Semester 2 Final Writing Exam – Sci/Med (2014/2015)

45

"My University Life"

I stady at , I'm medical student.
My stady is a very defficalt so I work hard. My subjets
are: English, Biostatestic, camistry, IT, Math and Bailogy.
Before university My daily routine is a butefull, but now
it's a very bad. I go to university at 6:00 a:m, but some
days I go to university at 10:00 a:m. IA 3days in a week
I have 4 classes and in 2 day in a week I have 2 classes. sometimes
I have lape classes, I eat my breakfast at 11:00 a:m, then I
don't eat anythings because I don't have anytimes. some days
I have full day. I'm back to my house at 3:00 p:m.
The studying is important for avryone. we need stadying
because we need a good carrer in the future. we need a good
life to live.

Figure 2a: Sample one of students' written texts from the end-of-year final exam



431 Semester 2 Final Writing Exam – Sci/Med (2014/2015)

Write at least 120 words for the writing task.

Make sure you write about all the parts of the writing task.

I study at many subjects that will benefit me later in life. These subjects include: Biology, Biostatistics, Physics, English, and Chemistry. Each subject is very important and benefitial in life. For example, studying English in prepatory year is crucial for a student who wants to go into medical school; because all the subjects are in English. Biology is also very important for me to learn, because it teaches me the basic science of organisms.

My life at includes many activities apart from learning benefitial subjects. It includes a daily routine that I never get bored from. First, I attend English class in the morning. Second, I go to a café with my friends to buy coffee and relax. Then, I attend the rest of my classes until 1:20 P.M. After that my friends and I go to pray. Finally, I attend my last class before going home. Some people may find my routine quite boring, but I love it!

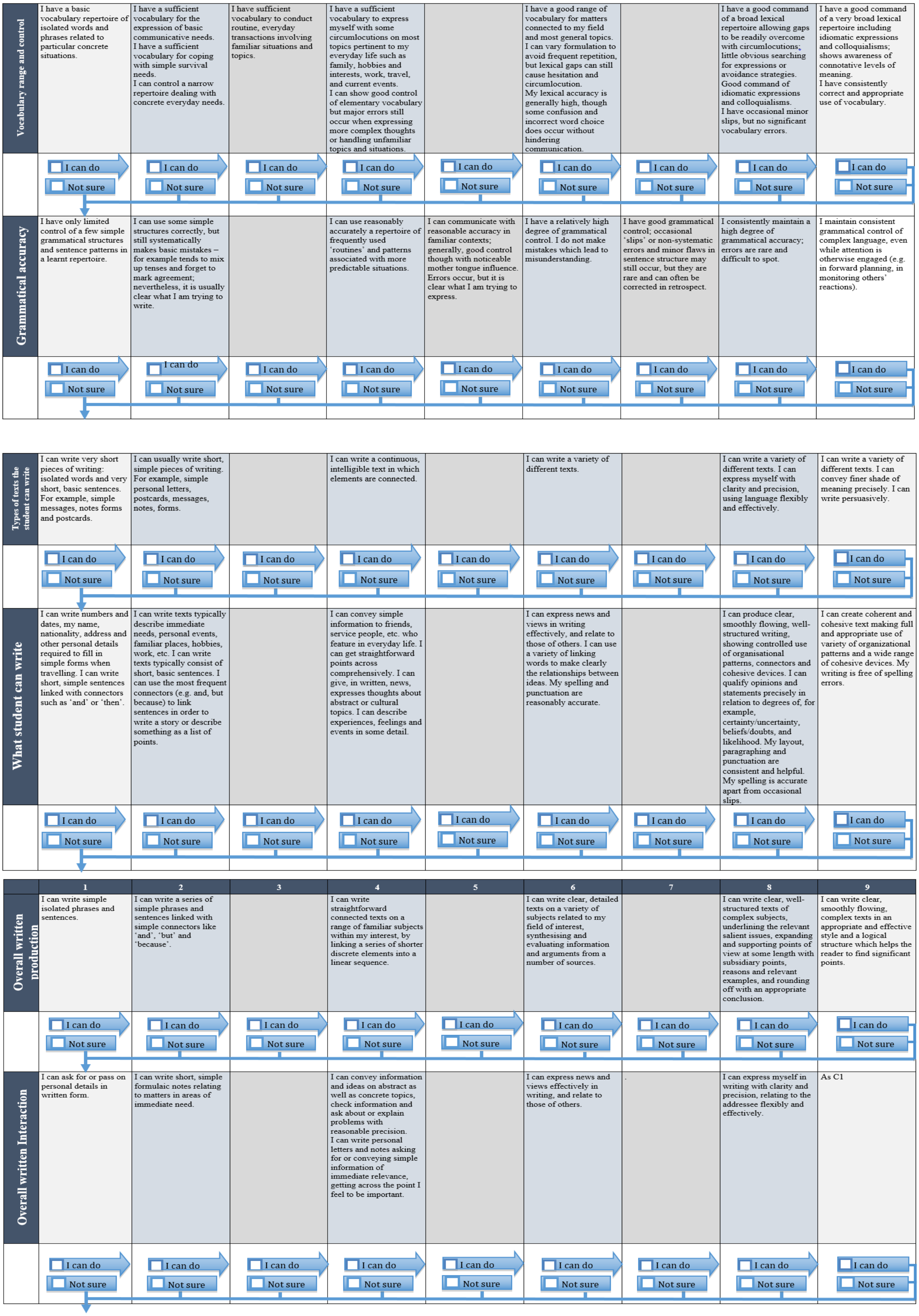
Studying at university is actually very important for many reasons. The first reason is that it raises a person's educational level. Second, it helps in getting a better job in the future. Last but not least, studying university improve's a person's social and academic skills.

Figure 2b: Sample two of students' written texts from the end-of-year final exam

8.2 Appendix 2. The student assessment grid

Reports and essays	No descriptors available. Start from number 4	No descriptors available. Start from number 4	No descriptors available. Start from number 4	I can write very brief reports to a standard conventionalized format, which pass on routine factual information and state reasons for actions.	I can write short, simple essays on topics of interest. I can summarize report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within my field with some confidence.	I can write an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. I can synthesise information and arguments from a number of sources.	I can write an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. I can evaluate different ideas or solutions to a problem.	I can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. I can expand and support points of view at some length with subsidiary points, reasons and relevant examples.	I can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literacy works. I can provide an appropriate and effective logical structure which helps the reader to find significant points.
	□ I can do □ Not sure	□ I can do □ Not sure	□ Can do □ Not sure	□ Can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure
Note Taking	No descriptors available. Please start reading from column number 4.	No descriptors available. Please start reading from column number 4.	No descriptors available. Please start reading from column number 4.	I can take notes as a list of key points during a straightforward lecture, provided the topic is familiar, and the talk is both formulated in simple language and delivered in clearly articulated standard speech.	I can take notes during a lecture which are precise enough for my own use at a later date, provided the topic is within my field of interest and the talk is clear and well-structured.	I can understand a clearly structured lecture on a familiar subject, and can take notes on points which strike me as important, even though I tend to concentrate on the words themselves and therefore to miss some information.		I can take detailed notes during a lecture on topics in my field of interest, recording the information so accurately and so close to the original that the notes could also be useful to other people.	I am aware of the implications and allusions of what is said and can make notes on them as well as on the actual words used by the speaker.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure

Orthographic control	I can copy familiar words and short phrases e.g. simple signs or instructions, names of everyday objects, names of shops and set phrases used regularly. I can spell my address, nationality and other personal details.	I can copy short sentences on everyday subjects – e.g. directions how to get somewhere. I can write with reasonable phonetic accuracy (but not necessarily fully standard spelling) short words that are in my oral vocabulary.		I can produce continuous writing which is generally intelligible throughout. Spelling, punctuation and layout are accurate enough to be followed most of the time.		I can produce clearly intelligible continuous writing which follows standard layout and paragraphing conventions. My spelling and punctuation are reasonably accurate but may show signs of mother tongue influence.		My layout, paragraphing and punctuation are consistent and helpful. My spelling is accurate, apart from occasional slips of the pen.	My writing is orthographically free of error.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure
Processing texts	I can copy out single words and short texts presented in standard printed format.	I can copy out short texts in printed or clearly handwritten format.	I can pick out and reproduce key words and phrases or short sentences from a short text within the learner's limited competence and experience.	I can collate short pieces of information from several sources and summarize them for somebody else. I can paraphrase short written passages in a simple fashion, using the original text wording and ordering.		I can summarize a wide range of factual and imaginative texts, commenting on and discussing contrasting points of view and the main themes. I can summarize extracts from news items, interviews or documentaries containing opinions, argument and discussion. I can summarize the plot and sequence of events in a film or play.		I can summarize long, demanding texts.	I can summarize information from different sources, reconstructing arguments and accounts in a coherent presentation of the overall result.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure



Vocabulary range and control	I have a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.	I have a sufficient vocabulary for the expression of basic communicative needs. I have a sufficient vocabulary for coping with simple survival needs. I can control a narrow repertoire dealing with concrete everyday needs.	I have sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.	I have a sufficient vocabulary to express myself with some circumlocutions on most topics pertinent to my everyday life such as family, hobbies and interests, work, travel, and current events. I can show good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.		I have a good range of vocabulary for matters connected to my field and most general topics. I can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. My lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.		I have a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms. I have occasional minor slips, but no significant vocabulary errors.	I have a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning. I have consistently correct and appropriate use of vocabulary.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure
Grammatical accuracy	I have only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.	I can use some simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what I am trying to write.		I can use reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	I can communicate with reasonable accuracy in familiar contexts; generally, good control though with noticeable mother tongue influence. Errors occur, but it is clear what I am trying to express.	I have a relatively high degree of grammatical control. I do not make mistakes which lead to misunderstanding.	I have good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	I consistently maintain a high degree of grammatical accuracy; errors are rare and difficult to spot.	I maintain consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure

Types of texts the student can write	I can write very short pieces of writing: isolated words and very short, basic sentences. For example, simple messages, notes forms and postcards.	I can usually write short, simple pieces of writing. For example, simple personal letters, postcards, messages, notes, forms.		I can write a continuous, intelligible text in which elements are connected.		I can write a variety of different texts.		I can write a variety of different texts. I can express myself with clarity and precision, using language flexibly and effectively.	I can write a variety of different texts. I can convey finer shade of meaning precisely. I can write persuasively.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure
What student can write	I can write numbers and dates, my name, nationality, address and other personal details required to fill in simple forms when travelling. I can write short, simple sentences linked with connectors such as 'and' or 'then'.	I can write texts typically describe immediate needs, personal events, familiar places, hobbies, work, etc. I can write texts typically consist of short, basic sentences. I can use the most frequent connectors (e.g. and, but because) to link sentences in order to write a story or describe something as a list of points.		I can convey simple information to friends, service people, etc. who feature in everyday life. I can get straightforward points across comprehensively. I can give, in written, news, expresses thoughts about abstract or cultural topics. I can describe experiences, feelings and events in some detail.		I can express news and views in writing effectively, and relate to those of others. I can use a variety of linking words to make clearly the relationships between ideas. My spelling and punctuation are reasonably accurate.		I can produce clear, smoothly flowing, well-structured writing, showing controlled use of organisational patterns, connectors and cohesive devices. I can qualify opinions and statements precisely in relation to degrees of, for example, certainty/uncertainty, beliefs/doubts, and likelihood. My layout, paragraphing and punctuation are consistent and helpful. My spelling is accurate apart from occasional slips.	I can create coherent and cohesive text making full and appropriate use of variety of organizational patterns and a wide range of cohesive devices. My writing is free of spelling errors.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure

	1	2	3	4	5	6	7	8	9
Overall written production	I can write simple isolated phrases and sentences.	I can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.		I can write straightforward connected texts on a range of familiar subjects within my interest, by linking a series of shorter discrete elements into a linear sequence.		I can write clear, detailed texts on a variety of subjects related to my field of interest, synthesising and evaluating information and arguments from a number of sources.		I can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.	I can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure
Overall written Interaction	I can ask for or pass on personal details in written form.	I can write short, simple formulaic notes relating to matters in areas of immediate need.		I can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision. I can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point I feel to be important.		I can express news and views effectively in writing, and relate to those of others.	.	I can express myself in writing with clarity and precision, relating to the addressee flexibly and effectively.	As C1
	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure	□ I can do □ Not sure

Figure 3 shows the 10 scales of the assessment grid that students were asked to complete.

8.3 Appendix 3.

Descriptive statistics for self-assessment, tutor assessment and rater scores

Tables 4 and 5 show the means and standard deviations for the scores for teacher- and rater-assessments, respectively.

Table 4. ***Descriptive analysis of PYP tutors' assessment across PYP levels***

	Elementary		Intermediate		Advanced	
	n=73		n=268		n=176	
CEFR Categories	M	SD	M	SD	M	SD
Overall Written Production	4.38	1.86	5.99	2.05	7.56	1.89
Overall Written Interaction	4.12	2.03	5.63	2.16	6.88	1.82
Type of Texts	4.46	2.24	5.80	2.21	7.28	1.80
What Can They Write	3.52	1.85	4.98	1.85	6.52	2.04
Vocabulary Range & Control	3.80	1.59	4.96	1.82	6.31	2.19
Grammatical Accuracy	3.88	1.89	4.98	1.74	6.16	2.24
Orthographic Control	4.22	2.52	4.89	1.83	6.97	1.88
Processing Texts	3.05	1.16	4.06	1.42	6.13	2.38
Reports and Essays	4.03	2.08	5.25	2.05	6.24	2.29
Note Taking	3.75	2.40	4.84	2.17	5.89	2.52
Average of Scales	3.79	1.45	5.12	1.60	6.65	1.54

M=Mean, SD=Standard deviation

Coding scheme for CEFR Categories: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)

Table 5. ***Descriptive analysis of the raters' assessment of sample students' texts across the PYP levels***

	Elementary		Intermediate		Advanced	
	n=14		n=55		n=36	
Rating Categories	M	SD	M	SD	M	SD
Range	3.57	1.21	3.90	1.32	5.05	1.28
Coherence	3.50	1.07	3.92	1.35	4.79	1.38
Accuracy	3.47	1.09	3.67	1.26	4.83	1.37
Description	3.55	1.22	3.82	1.28	4.86	1.36
Overall	3.56	1.13	3.87	1.29	4.96	1.28
Average score	3.53	1.14	3.83	1.30	4.88	1.33

M=Mean, SD=Standard deviation

Coding scheme for Manual Grid: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)

8.4 Appendix 4.

Differences between elementary, intermediate and advanced groups on students' self-assessments

One-way ANOVA was used to identify differences across the PYP levels for the students' assessments. After performing the analysis, Levene's test (Levene 1960) was checked. This test "tests whether the variance in scores is the same for each of the three groups" (Pallant 2013: 262). Where Levene's test indicated there was no violation of the assumption of homogeneity of variance, ANOVA was used (Table 4); when the assumption of equal variances was violated, the non-parametric analysis of variance (Brown-Forsythe and Welch Tests), as mentioned in Green (2013), were used instead (Table 5).

If the significance (P-value) was <0.05, this indicates a significant difference between the mean scores between the three groups. However, this does not show "which group is different from which other group" (Pallant 2013: 262). For this reason, a post-hoc test, i.e., Tukey's Honestly Significant Difference (HSD) test (Pallant 2013) (if there is no violation to the assumption of homogeneity of variances; Table 6) or Tamhane's T2 (Green 2013) (with heterogeneity of variances; Table 7), were used to check the significance between each pair of the three PYP groups. Post-hoc tests are only utilised if significant differences in means are identified (Pallant 2013: 263).

Table 6 shows the CEFR-based categories for which ANOVA was used.

Table 6. ***One-way analysis of variance of students' self-assessment of CEFR levels across PYP levels***

CEFR Categories	*SS*	*df*	*MS*	*F*	*P-value*	*η²*
What Students Can Write						
Between Groups	488.83	2	244.42	52.58	<0.001	0.16
Within Group	2393.82	515	4.65			
Total	2882.65	517				
Reports and Essays						
Between Groups	634.05	2	317.02	60.31	<0.001	0.19
Within Group	2686.11	511	5.26			
Total	3320.16	513				
Note Taking						
Between Groups	279.96	2	139.98	26.89	<0.001	0.095
Within Group	2665.77	512	5.21			
Total	2945.74	514				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η^2=Effect size M=Mean, SD=Standard deviation, df=degrees of freedom, η^2=Effect size: 0.02=small; 0.13=medium; 0.26=large.

Table 7 shows tests for equality of means for which non-parametric tests were used.

Table 7. ***Robust test of equality of mean of students' self-assessment of their CEFR levels across the three PYP levels***

CEFR Categories	*Statistic*	*df1*	*df2*	*P-value*
Overall Written Production				
Welch	56.05	2	186.89	<0.001
Brown-Forsythe	46.07	2	219.18	<0.001
Overall Written Interaction				
Welch	61.47	2	199.63	<0.001
Brown-Forsythe	69.48	2	338.76	<0.001

CEFR Categories	*Statistic*	*df1*	*df2*	*P-value*
Type of Texts				
Welch	44.49	2	199.82	<0.001
Brown-Forsythe	49.86	2	338.40	<0.001
Vocabulary Range & Control				
Welch	46.06	2	194.25	<0.001
Brown-Forsythe	51.53	2	316.85	<0.001
Grammatical Accuracy				
Welch	13.99	2	188.51	<0.001
Brown-Forsythe	14.90	2	282.66	<0.001
Orthographic Control				
Welch	29.50	2	191.96	<0.001
Brown-Forsythe	25.11	2	242.60	<0.001
Processing Texts				
Welch	52.33	2	205.86	<0.001
Brown-Forsythe	62.06	2	362.55	<0.001

df=degrees of freedom

Table 8 shows the post hoc Tukey honestly significant difference (HSD) test of pairwise differences between groups on student self-assessments.

Table 8. ***Post-hoc Tukey HSD of students' self-assessment of their CEFR levels across the three PYP levels (for items with homogeneity of variances)***

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value
What Students Can Write	Elementary	Intermediate	-0.48	0.29	0.22
		Advanced	**-2.40***	0.30	<.001
	Intermediate	Advanced	**-1.92***	0.21	<0.001
Reports and Essays	Elementary	Intermediate	-0.36	0.31	0.46
		Advanced	**-2.61***	0.32	<0.001
	Intermediate	Advanced	**-2.25***	0.22	<0.001
Note Taking	Elementary	Intermediate	-0.22	0.30	0.74
		Advanced	**-1.72***	0.32	<0.001
	Intermediate	Advanced	**-1.50***	0.22	<0.001
Conditions and Limitations	Elementary	Intermediate	-0.45	0.44	0.57
		Advanced	**-2.19***	0.46	<0.001
	Intermediate	Advanced	**-1.74***	0.31	<0.001

Table 9 shows the post hoc Tamhane test of pairwise differences between groups on student self-assessments for items with heterogeneity of variances.

Table 9. *Post hoc Tamhane test (heterogeneity of variances) of students' self-assessment of their CEFR levels across the three PYP levels*

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value
Overall Written Production	Elementary	Intermediate	-0.67	0.31	0.092
		Advanced	**-2.34***	0.30	<0.001
	Intermediate	Advanced	**-1.67***	0.18	<0.001
Overall Written Interaction	Elementary	Intermediate	-0.30	0.28	0.66
		Advanced	**-2.74***	0.31	<0.001
	Intermediate	Advanced	**-2.44***	0.24	<0.001
Types of Texts the Students can write	Elementary	Intermediate	-0.33	0.28	0.55
		Advanced	**-2.33***	0.31	<0.001
	Intermediate	Advanced	**-2.00***	0.23	<0.001
Vocabulary Range & Control	Elementary	Intermediate	-0.40	0.26	0.34
		Advanced	**-2.30***	0.30	<0.001
	Intermediate	Advanced	**-1.90***	0.22	<0.001
Grammatical Accuracy	Elementary	Intermediate	-0.77	0.35	0.083
		Advanced	**-1.81***	0.38	<0.001
	Intermediate	Advanced	**-1.04***	0.26	<0.001
Orthographic Control	Elementary	Intermediate	-0.36	0.36	0.70
		Advanced	**-1.93***	0.36	<0.001
	Intermediate	Advanced	**-1.57***	0.23	<0.001
Processing Texts	Elementary	Intermediate	**-.58***	0.21	0.020
		Advanced	**-2.31***	0.25	<0.001
	Intermediate	Advanced	**-1.73***	0.20	<0.001

Bold with *=significant results

8.5 Appendix 5

Differences between elementary, intermediate and advanced groups for tutor assessments

One-way ANOVA was used to identify differences across the PYP levels for the tutor assessments. Where there was no violation of the assumption of homogeneity of variance, ANOVA was used (Table 10); when the assumption of equal variances was violated, the non-parametric analysis of variance (Brown-Forsythe and Welch Tests) were used (Table 11). A post-hoc Tukey's HSD (if there is no violation to the assumption of homogeneity of variances; Table 12) or Tamhane's T2 (with heterogeneity of variances; Table 13) were used.

Table 10. *One-way analysis of variance of tutors' assessment across PYP levels*

CEFR Categories	*SS*	*df*	*MS*	*F*	*P-value*	*η²*
Overall written Production						
Between Groups	654.09	2	327.05	84.91	<0.001	0.24
Within Group	2006.79	521	3.85			
Total	2660.88	523				
What Students Can Write						
Between Groups	590.42	2	295.21	80.05	<0.001	0.23
Within Group	1928.81	523	3.69			
Total	2519.22	525				
Reports and Essays						
Between Groups	253.37	2	126.69	27.38	<0.001	0.10
Within Group	2221.15	480	4.63			
Total	2474.52	482				
Note Taking						
Between Groups	250.54	2	125.27	22.78	<0.001	0.08
Within Group	2640.20	480	5.50			
Total	2890.74	482				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η²=Effect size: 0.02=small; 0.13=medium; 0.26=large.

Table 11. *Robust test of equality of mean of tutors' assessment across PYP levels*

CEFR Scales	*Statistic*	*df1*	*df2*	*P-value*
Overall Written Interaction				
Welch	63.84	2	242.69	<0.001
Brown-Forsythe	60.26	2	357.31	<0.001
Type of Texts				
Welch	64.67	2	235.86	<0.001
Brown-Forsythe	59.86	2	317.83	<0.001
Vocabulary Range & Control				
Welch	59.00	2	253.36	<0.001
Brown-Forsythe	60.82	2	426.13	<0.001
Grammatical Accuracy				
Welch	40.63	2	233.06	<0.001
Brown-Forsythe	44.31	2	366.37	<0.001
Orthographic Control				
Welch	77.17	2	159.67	<0.001
Brown-Forsythe	63.16	2	166.48	<0.001
Processing Texts				
Welch	94.79	2	194.58	<0.001
Brown-Forsythe	116.16	2	357.74	<0.001

df=degrees of freedom

Table 12. *Tukey HSD of tutors' assessment across the three PYP levels*

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value
Overall Written Production	Elementary	Intermediate	**-1.61***	0.24	<0.001
		Advanced	**-3.18***	0.25	<0.001
	Intermediate	Advanced	**-1.57***	0.19	<0.001
Types of Texts Students can write	Elementary	Intermediate	**-1.35***	0.26	<0.001
		Advanced	**-2.82***	0.27	<0.001
	Intermediate	Advanced	**-1.48***	0.20	<0.001
Reports and Essays	Elementary	Intermediate	**-1.22***	0.31	<0.001
		Advanced	**-2.21***	0.31	<0.001
	Intermediate	Advanced	**-0.99***	0.21	<0.001
Note Taking	Elementary	Intermediate	**-1.10***	0.33	<0.001
		Advanced	**-2.15***	0.34	<0.001
	Intermediate	Advanced	**-1.05***	0.23	<0.001
Average of all scales	Elementary	Intermediate	**-1.33***	0.19	<0.001
		Advanced	**-2.86***	0.20	<0.001
	Intermediate	Advanced	**-1.53***	0.15	<0.001

Table 13. *Post hoc Tamhane of tutors' assessment across the three PYP levels*

Dependent Variable	(I) PYP levels	(J) PYP levels	Mean difference (I-J)	Std. error	P-value	95% Confidence interval	
						Lower bound	Upper bound
Overall Written Interaction	Elementary	Intermediate	**-1.51***	0.25	<0.001	-2.12	-.89
		Advanced	**-2.76***	0.25	<0.001	-3.37	-2.15
	Intermediate	Advanced	**-1.26***	0.19	<0.001	-1.71	-.80
What students Can Write	Elementary	Intermediate	**-1.46***	0.23	<0.001	-2.01	-.91
		Advanced	**-2.99***	0.25	<0.001	-3.58	-2.40
	Intermediate	Advanced	**-1.54***	0.19	<0.001	-1.99	-1.08
Vocabulary Range and Control	Elementary	Intermediate	**-1.16***	0.20	<0.001	-1.65	-.67
		Advanced	**-2.52***	0.23	<0.001	-3.06	-1.95
	Intermediate	Advanced	**-1.35***	0.20	<0.001	-1.82	-.88
Grammatical Accuracy	Elementary	Intermediate	**-1.11***	0.23	<0.001	-1.66	-.56
		Advanced	**-2.29***	0.26	<0.001	-2.91	-1.67
	Intermediate	Advanced	**-1.18***	0.20	<0.001	-1.65	-.71
Orthographic Control	Elementary	Intermediate	**-0.67**	0.34	0.147	-1.50	.16
		Advanced	**-2.75***	0.35	<0.001	-3.59	-1.91
	Intermediate	Advanced	**-2.08***	0.18	<0.001	-2.51	-1.65
Processing Texts	Elementary	Intermediate	**-1.51***	0.25	<0.001	-1.43	-.59
		Advanced	**-2.76***	0.25	<0.001	-3.63	-2.54
	Intermediate	Advanced	**-1.26***	0.19	<0.001	-2.54	-1.60

Bold with *=significant results

8.6 Appendix 6

Differences between elementary, intermediate and advanced groups on rater assessments

Table 14 shows the ANOVA for differences across the PYP levels for the rater assessments and Table 15 shows the post-hoc Tukey's HSD.

Table 14. ***One way ANOVA of raters' assessment across PYP levels***

		SS	*df*	*MS*	*F*	*P-value*	*η²*
	Between Groups	37.28	2	18.64			
Range	Within Groups	66.32	99	0.67	27.823	*p<0.001*	0.36
	Total	103.59	101				
	Between Groups	24.64	2	12.32			
Coherence	Within Groups	65.04	99	0.66	18.76	*p<0.001*	0.28
	Total	89.7	101				
	Between Groups	35.33	2	17.66			
Accuracy	Within Groups	60.32	99	0.61	28.99	*p<0.001*	0.37
	Total	95.65	101				
	Between Groups	29.93	2	14.97			
Description	Within Groups	61.04	99	0.62	24.28	*p<0.001*	0.33
	Total	90.97	101				
	Between Groups	33.23	2	16.61			
Overall	Within Groups	64.11	99	0.65	25.66	*p<0.001*	0.34
	Total	97.34	101				

SS=Sum of squares, df=degrees of freedom, MS=mean square, F=F ratio, η^2=Effect size: 0.02=small; 0.13=medium; 0.26=large.

Table 15. *Post hoc Tukey analysis of range, coherence, accuracy, description, and overall grouped by PYP levels*

Dependent Variable	(I) PYP Levels	(J) PYP Levels	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
						Lower Bound	Upper Bound
Range	Elementary	Intermediate	-0.33	0.25	0.40	-0.93	0.27
		Advanced	-1.53	0.27	**<0.001***	-2.15	-0.89
	Intermediate	Advanced	-1.19	0.18	**<0.001***	-1.61	-0.77
Coherence	Elementary	Intermediate	-0.42	0.25	0.227	-1.01	0.18
		Advanced	-1.33	0.26	**<0.001***	-1.97	-0.70
	Intermediate	Advanced	-.92	0.18	**<0.001***	-1.33	-0.50
Accuracy	Elementary	Intermediate	-0.20	0.24	0.674	-0.78	0.37
		Advanced	-1.40	0.25	**<0.001***	-2.00	-0.79
	Intermediate	Advanced	-1.19	0.17	**<0.001***	-1.60	-0.79
Description	Elementary	Intermediate	-0.27	0.24	.0507	-0.85	0.31
		Advanced	-1.34	0.26	**<0.001***	-1.95	-0.74
	Intermediate	Advanced	-1.07	0.17	**<0.001***	-1.48	-0.67
Overall	Elementary	Intermediate	-0.31	0.25	0.423	-0.90	0.28
		Advanced	-1.44	0.26	**<0.001***	-2.06	-0.81
	Intermediate	Advanced	-1.12	0.17	**<0.001***	-1.54	-0.71

Bold with *=significant results

8.7 Appendix 7

RQ2: Student versus teachers paired t-test and correlation

Table 16 shows the paired t-test between students and teachers for each scale, separated by PYP level.

Table 16. ***Paired differences between self-and tutors' assessment in each PYP level***

CEFR Scales	PYP students		PYP tutors					Cohen's
	M	SD	M	SD	t	df	*P*	d
Elementary (n=72)								
Overall Written Production	5.62	2.33	4.41	1.92	3.72	70	**<0.001**	0.44
Overall Written Interaction	3.96	2.10	4.11	2.12	-0.56	70	0.576	-0.07
Type of Texts	3.94	2.06	4.46	2.38	-1.54	70	0.128	-0.18
What Can They Write	4.40	2.26	3.47	1.85	3.07	71	**0.003**	0.36
Vocabulary Range & Control	3.59	2.00	3.86	1.66	-0.91	70	0.367	-0.11
Grammatical Accuracy	4.34	2.70	3.85	1.95	1.41	70	0.164	0.17
Orthographic Control	4.78	2.92	4.24	2.59	1.04	50	0.304	0.15
Processing Texts	3.80	1.61	3.00	1.22	2.84	50	**0.006**	0.40
Reports and Essays	4.00	2.62	4.10	2.05	-0.22	49	0.826	-0.03
Note Taking	5.04	2.69	3.80	2.46	2.70	50	**0.009**	0.38
Average Scales	4.49	1.59	3.97	1.65	2.24	71	**0.028**	0.26
Intermediate (n=232)								
Overall Written Production	6.26	2.17	5.97	2.08	1.52	226	0.129	0.10
Overall Written Interaction	4.23	2.33	5.60	2.17	-6.61	226	**<0.001**	-0.44
Type of Texts	4.28	2.26	5.79	2.25	-7.77	228	**<0.001**	-0.51
What Can They Write	4.78	2.25	4.94	1.86	-0.85	228	0.394	-0.06
Vocabulary Range & Control	3.87	1.95	4.94	1.86	-6.64	230	**<0.001**	-0.44
Grammatical Accuracy	5.05	2.37	4.95	1.75	0.57	230	0.570	0.04
Orthographic Control	5.47	2.70	4.87	1.86	2.88	217	**0.004**	0.19
Processing Texts	4.36	1.70	4.01	1.41	2.48	217	2.014	0.17
Reports and Essays	4.55	2.31	5.20	2.10	-3.09	210	**0.002**	-0.21
Note Taking	5.43	2.18	4.81	2.20	3.00	211	**0.003**	0.21
Average Scales	4.89	1.51	5.18	1.67	-2.15	230	**0.032**	-0.14
Advanced (n=170)								
Overall Written Production	7.96	1.65	7.62	1.82	1.87	168	0.064	0.14
Overall Written Interaction	6.74	2.56	6.90	1.83	-0.66	168	0.510	-0.05
Type of Texts	6.35	2.47	7.24	1.81	-4.01	169	**<0.001**	-0.31
What Can They Write	6.86	1.95	6.56	2.05	1.45	169	0.150	0.11
Vocabulary Range & Control	5.86	2.40	6.31	2.19	-1.75	168	0.082	-0.13
Grammatical Accuracy	6.14	2.89	6.19	2.19	-0.19	168	0.847	-0.01
Orthographic Control	6.99	2.16	7.05	1.78	-0.30	168	0.762	-0.02
Processing Texts	6.12	2.22	6.24	2.34	-0.51	168	0.613	-0.04
Reports and Essays	6.78	2.07	6.26	2.26	2.18	169	**0.030**	0.17
Note Taking	6.90	2.19	5.96	2.43	3.84	168	**<0.001**	0.30

M= Mean, SD=Standard deviation

Coding scheme for CERF Scales: 1 (A1); 2 (A2); 3 (A2+); 4 (B1), 5 (B1+); 6 (B2); 7 (B2+); 8 (C1); 9 (C2)

Cohen’s d_z calculated as Mean misalignment/SD of misalignment. Cohen’s d calculated as 2 x t/sqrt, 0.2=small effect; 0.5=medium; 0.8=large

Bold = significant result

Table 17 shows the correlation between students and teachers’ scores, the weighted kappa (measure of agreement) and the percentages of scores with exact agreement (identical level assigned), or agreements within one or two levels.

Table 17. ***Correlation and agreement between ratings of self- and tutors’ assessment***

CEFR Scales	Correlation (r) (n=517)	Weighted Kappa (n=517)	% exact agreement	% within one adjacent CEFR level	% within two adjacent CEFR levels
Overall Written Production	0.29 P<0.001	0.27	31.5	38.9	65.5
Overall Written Interaction	0.22 P<0.001	0.22	23.3	33.2	62.7
Types of Texts the Students can write	0.29 P<0.001	0.25	23.6	31.5	60.4
What Students can write	0.28 P<0.001	0.28	25.7	31.6	67.9
Vocabulary Range and Control	0.25 P<0.001	0.25	21.7	35.2	61.6
Grammatical Accuracy	0.23 P<0.001	0.19	15.9	40.8	61.8
Orthographic Control	0.26 P<0.001	0.26	21.5	31.3	68.0
Processing Texts	0.30 P<0.001	0.32	29.9	48.4	73.7
Reports and Essays	0.23 P<0.001	0.15	20.2	45.9	65.0
Note Taking	0.18 P<0.001	0.15	22.7	39.4	59.5

8.8 Appendix 8.

RQ2. Comparisons of students, teachers and raters' assessments

Table 18 and 19 show the Tukey's post hoc tests, firstly (Table 16) with data for all students across the PYP levels and secondly (Table 17) separated by PYP level.

Table 18. *Tukeys post hoc analysis for scores grouped as to the type of raters*

(I) Type	(J) Type					
	Students' self-assessment		Teachers' assessment		Raters' assessment	
	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value
Students' self-assessment			-0.031	0.99	1.28*	<0.001
Teachers' assessment	0.031	0.99			1.31*	<0.001
Raters' assessment	-1.28*	<.001	-1.31*	<0.001		

* The mean difference is significant at the 0.05 level.

Table 19. *Post hoc Tukey analysis of PYP level grouped by assessor and level*

PYP Levels	(I) Type	(J) Type	Mean Difference (I-J)	p-value
Elementary	Self	Tutors	-0.03	0.998
		Raters	1.44*	0.004
	Tutors	Raters	1.47*	0.004
Intermediate	Self	Tutors	-0.22	0.67
		Raters	0.92*	0.001
	Tutors	Raters	1.14*	<0.001
Advanced	Self	Tutors	0.26	0.67
		Raters	1.77*	<0.001
	Tutors	Raters	1.52*	<0.001

* The mean difference is significant at the 0.05 level.