Journal of Language Horizons

Scientefic Quarterly Journal of Language Horizons, *Alzahra University*
Volume 7, Issue 1, Spring 2023 (Serial No. 15)

ESL Writers' Performance in Exam and Non-Exam Academic Writing Settings

Research Article
pp. 59-85

Fahimeh Marefat[*1]

Mojtaba Heydari[2]

Received: 2021/08/29 Accepted: 2022/02/12

Abstract

With the growing access to new types of reference tools, today's L2 writers have a plethora of choices when completing an academic writing assignment. Such resources are absent in most high-stakes academic writing exams, making the two situations dissimilar. Aimed to compare the performances of ESL writers in Exam and Non-exam (real-life) academic writing situations, the present study recruited seven ESL university students who had previously taken an IELTS test. The students completed two analogous writing tasks: an exam-setting and a Non-exam writing test which aimed to simulate the real-life setting. Coh-Metrix analysis of the linguistic features of syntactic complexity, lexical sophistication, and text cohesion of the writings suggested that the students improved the textual quality of their writings in real-life academic writing situation. In addition, FACETS analysis of the quality of the writings, as assessed by the human raters, showed that the students did not benefit equally from the merits of the real-life settings compared to the Exam settings. The findings suggest that the students spent different amounts of time and used different types of queries to consult with external resources. Students' background training and writing strategies can highly affect their performance in real-life academic writing compared to the writing exams, warning against the validity of such tests.

Keywords: academic writing, Coh-Metrix, exam / non-exam settings, FACETS, writing exams validity

---

[*] Corresponding Author
[1] Professor, Allameh Tabataba'i University, Department of English Language and Literature, Faculty of Persian Literature and Foreign Languages, Tehran, Iran. fmarefat@atu.ac.ir. fahimehmarefat@yahoo.com
[2] Ph.D. Candidate, Allameh Tabataba'i University. Department of English Language and Literature, Faculty of Persian Literature and Foreign Languages, Tehran, Iran.Heydari_mojtaba@atu.ac.ir

DOI: 10.22051/lghor.2022.37489.1553
DOR: 20.1001.1.2588350.2023.7.1.3.1

Introduction

In the digital era, students' daily lives have been occupied by a variety of technology-facilitated practices such as web-browsing, texting, emailing, playing online games, and chatting through social network platforms (Zheng & Warschauer, 2017). In addition to their general life practices, students' academic lives heavily rely on digital resources as well: taking notes in the class, doing research, reading e-books, doing and submitting their course assignments, and being in contact with their teachers and peers. Regarding academic writing, a myriad of digital tools is available in assisting learners to compose their essays and enormously facilitating their process of writing. Unsurprisingly, today's students do most, if not the entire, of their academic writings using computers and, as a result, have a wider choice of tools to compose, edit, and enrich their writings (Yoon, 2016; Zhi & Huang, 2021).

Dictionaries and other traditional writing references are now available online and provide richer, faster, and easier access to information. (Dziemianko, 2012). Moreover, new types of online resources, like search engines, grammar checkers, corpus tools, and forums are now helping L2 writers to solve their lexical and grammatical problems (Yoon, 2016). Due to the significance of digital resources in learners' life, the ability to effectively utilize different reference resources should shape an essential part of digital literacy in academic settings (Conroy, 2010; Flowerdew, 2010; Kennedy & Miceli, 2010). Similarly, the use of online information is becoming an indispensable element of writing (Leijten et al., 2014). Therefore, as Hayes (2012) discusses, it should be included in future research to help achieve a better understanding of real-world L2 writing behaviors (Gánem-Gutiérrez & Gilmore, 2018).

However, a strong disconnect exists between students' classroom writing and exam writing contexts. While students are benefiting from the countless number of digital tools which assist them in the process of everyday classroom writing, the role of such resources is utterly overlooked in most high-stakes English exams. Today, while in many popular high-stakes academic writing tests such as IELTS (computer-delivered) and TOEFL, computers are the medium of text composition, the only digital asset at the writers' disposal is the text processor exclusively developed for composing the text. Therefore, students who have been practicing writing by getting assistance from digital resources, are deprived of such tools and

solely rely on their mental resources to write an essay.

Writing Processes

In the last two decades, the distinction between planning, composing, and revising has begun to erode entirely as a writer's craft has shifted from taking notes on papers to type and even dictate them (e.g., by Google, Siri, etc.) to be stored on their mobiles, laptops, or other digital companions. While the older models of writing emphasized the role of writer's memory in writing, recent models of writing additionally include the element of searching which recognizes the writer's use of external sources, like online dictionaries, to access information during writing (Leijten et al., 2014). The cognitive-based writing models consider writing as a problem-solving activity where the writers should approach the task as a problem and employ intellectual resources to solve it (Hyland, 2002). Flower and Hayes' (1981) model, for example, suggested that the writing process is recursive and involves planning, drafting, revising, and editing.

Inasmuch as writing processes have evolved in the last decades, the methods to observe and study these processes, too, have changed. L2 writing researchers have traditionally studied composition processes using direct observation of writers and their reflections and recounts on their writings to explore L2 writers' problem-solving processes and strategies (Bloom, 2008; Serror, 2013). With the emergence and development of digital technology, revolutionary observation methods including screen capture (e.g., Khuder & Harwood, 2015), keystroke logging (e.g., Serror, 2013), and eye-tracking (e.g., Gánem-Gutiérrez & Gilmore, 2018) have brought up more in-depth yet fairly quantifiable data to study writing composition processes.

Previous Studies

Despite the burgeoning use of digital resources, research on learner use of online resources has mainly been limited to individual reference resources in classroom settings (Yoon, 2016). Besides, studies comparing the writing processes and performances in exam and non-exam situations are scarce. Roca de Larios et al. (2008), for instance, investigated how different SL writers allocated time in the process of writing. They came up with two main conclusions: (i) the largest

percentage of composition time was spent on formulation and it was the predominant process for all groups and (ii) writers with different proficiency levels devoted different proportions of time to the writing processes; more skilled writers were more likely to regulate their composition processes.

In a recent study concerning the time of writing, Lee et al. (2021) found that when the L2 writers were given extra 15 minutes in a 30-minute argumentative writing test, the quality of argumentation was significantly higher in all the related subscores. In a very relevant study, Oh (2019) gave 39 English learners the chance to use extra resources to write an online review and compared their performance when they were not allowed to use the writing resources. He found that although the students performed better when they used resources during the writing process, their relative proficiency standings did not change. Oh concluded that giving access to the writing resources did not show any difference in distinguishing test-takers' levels.

Also, Khuder and Harwood (2015) investigated the product and process of writing in different situations. Ten graduate students wrote two argumentative essays under test and non-test conditions. The researchers used keystroke logging, screen recording, and stimulated recall protocols to observe their writing processes. They found statistically different time allocation for writing processes under the two settings. Besides, the participants received an average of .8 points higher for writing under the non-test condition. However, in Khuder and Harwood's study, the non-test situation failed to represent an authentic situation where students usually write the writing assignments. For instance, one of the researchers was present as the participants were writing under the Non-exam situation. In addition, as reported by the researchers, none of the participants, except for one, spent a significantly longer time in the non-test situation. This casts doubt on whether the students felt at ease, similar to a real-life situation, to spend as much time as they needed to complete their writing. Still more, authentic writing requires more than one sitting (Gánem-Gutiérrez & Gilmore, 2018).

This Study

As part of a bigger research project, this study set out to investigate how L2 writers navigate the growing variety of digital resources while completing an academic writing assignment and compare it against how they complete the task in a

situation similar, in many ways, to exam settings. This will provide insights as to how different learners can benefit from digital resources and how such affordances can reshape the future of L2 exams.

Research Questions

The present study sought the answer to the following research questions:

a) Is there any difference in the ESL writers' performance in terms of linguistic features, and overall, under the two academic writing settings of Exam vs. Non-exam?
b) Does the provision of an extended time for the writing process affect student's writing equally?
c) Does the provision of extra resources (e.g., dictionaries, wikis, etc.) affect student's writing equally?

Method

Participants

The call for participation was announced in five different ESL classes in a Canadian university with international students from a variety of language backgrounds. The participants were invited to take part in the research, being informed about the nature of the study and the time required to accomplish the tasks. From among the 83 respondents who expressed their initial willingness by filling out the online survey form, seven participants managed to finish the research procedure by attending all three phases of the study. Although this number of participants may be insufficient for quantitative studies, due to the qualitative nature of data collection and analysis in writing process studies (e.g., Khuder & Harwood, 2015) and the time and resources required to collect data for each participant, it seemed adequate for the purpose and scope of our study.

The students were from five language backgrounds (Persian, Spanish, Chinese, Arabic, and French) studying at graduate and undergraduate programs studying Engineering, Management, Linguistics, Philosophy, and Communication. At the time of the study, all the participants (four male and three were female students between 19 and 35) had an IELTS score (within two years of the time of the test). Also, three of them were preparing for a higher score. Table 1 shows the participants' biodata.

Table 1

The Participants' Biodata

Name	Sex	Age	L1	Major	School Year	IELTS Writing Band Score
Abdul	Male	35	Arabic	Linguistics	Post-graduate	7.5
Ali	Male	32	Persian	Management	Post-graduate	6.5
Lio	Male	19	Chinese	Engineering	Undergraduate	5.5
Pablo	Male	20	Spanish	Engineering	Undergraduate	6
Pari	Female	31	Persian	Engineering	Post-graduate	6.5
Sophie	Female	24	French	Philosophy	Graduate	7
Wei	Female	26	Chinese	Management	Graduate	6.5

Instruments

Computer Familiarity Questionnaire (CFQ). As suggested by previous research, the participants' familiarity with computers might have an impact on their performance in this study. Therefore, a CFQ was adapted from Weir et al.'s (2007) study, with a few adjustments for the present research context. The CFQ (see Appendix A) consists of 14 questions on a 5-point Likert Scale. A link to the online version of the questionnaire was sent to the email addresses of the participants. The CFQ took about 10 minutes on average for each participant to complete. The analysis of the results pointed out that the participant's familiarity with computers in terms of three aspects, i.e. Computer Usage, Comfort and Perceived Ability, and Interest in Computers were very high for all the participants as they all reported using computers very frequently at home and university.

The Writing Tasks. Two writing tasks were designed to collect student writings in the exam and real-life settings. In the Exam-setting, the participants were asked to write an essay in a 40-minute time limit on argumentative writing prompt with a general academic topic (Appendix B). While the writers used a computer to write their essays, they were only allowed to use Notepad to type their texts. This way, they didn't benefit from any spell/grammar-checkers, dictionaries, or thesauruses that might be enabled in a more sophisticated word processing software like MS Word. To simulate a high-stake writing exam setting, they were not allowed to use the internet, books, or any software during the 40-minute writing session.

In the Non-exam setting, the students were again asked to write an argumentative essay in response to a prompt similar to the one assigned to them in Exam-setting (Appendix B). However, in this task, the students were not given a 40-minute time limit. Instead, a 10-day deadline was provided to write their essays with no limitation in the hours or the number of sessions spent on the writing task within the deadline time frame. To make the task similar to the real-life academic writing tasks, the writers were given the choice to take our laptop equipped with the screen recording software to any preferred place (e.g., home, library), during the 10-day deadline to finish and submit their texts. During that time, they were also allowed to consult with any online resources including dictionaries, grammar-checkers, wikis, etc. To ensure maximum similarity between the two prompts, both prompts were selected from a pool of practice IELTS writing prompts provided by the British Council website. As for the comparability of the two prompts and the essays, two experienced IELTS examiners were consulted.

Screen Recording Software. To record the screen activities of the writers during both tasks, ApowerREC v 1.5 was installed on our laptop. The tool allows the users to easily save the screen-captured video and share it with others. We decided not to use keystroke logging since the writers spent some of their time using external resources such as Internet browsing which is not captured in keystroke logging software like Inputlog (Khuder and Harwood, 2015). Instead, we relied on the screen capture video to manually code the episodes of the writing, despite being more time-consuming, to capture and code all the elements of using external resources like searching and reading.

The Raters. Three raters were asked to blindly evaluate the essays written by the participants in the two settings. The raters were not informed that the texts were written by the same individuals under Exam and Non-exam settings. The raters were all IELTS teachers who had extensive experience in grading IELTS writing grading and were working as 'IELTS Mock Examiners' at language institutes. The raters were all males, at the ages of 33, 35, and 45 years old. The raters used the public version of the IELTS writing rubric to grade the tests and employed a 0-9 band score range based on the rubric.

Coh-Metrix Indices. One of the purposes of this study was to assess the degree to which linguistic features of the essays crafted in the Exam and Non-exam

settings differ. To compare the linguistic features of syntactic complexity, lexical sophistication, and text cohesion, the automatic computational linguistic tool of Coh-Metrix was employed (McNamara & Graesser, 2012).

Several indices were drawn from Coh-Metrix 3.0 to assess the quality of the essays across word, sentence, and discourse levels. The indices were selected based on the prior research and their application in the present study. For example, McNamara et al. (2010) found that syntactic complexity, lexical diversity, and word frequency were better predictors of essay quality. In the current study, we came up with 11 Coh-Metrix indices which proved valid in the literature and matched the IELTS writing assessment criteria. In addition, as accuracy is not measured by Coh-Metrix, we included one index to assess the accuracy of the essays. Also, one category was added to measure the length of words and sentences. The indices were categorized into five criteria and included:

a) *Length:* word count, word length, and sentence length
b) *Lexical Complexity:* word familiarity, lexical diversity, word frequency, content word concreteness
c) *Syntactic Complexity:* embedded clauses, number of modifiers per noun phrase, syntactic similarity, minimal edit distance
d) *Cohesion:* aspect repetition, content word overlap, connective incidences
e) *Accuracy:* error-free T-units divided by total T-units

Procedure

Phase 1. Exam-Setting. This phase aimed to simulate IELTS taks 2 writing tests. The writers were instructed to write no less than 250 words in 40 minutes in response to the selected prompt. The students completed this phase in a quiet university office using our equipped laptop. To collect the data for later analysis, ApowerREC software was used to record the screen activity of the writers from the beginning to the end of the task. At the end of the session, the writing document was saved in .txt format for later analysis.

Phase 2. Non-exam Setting. This phase was designed to simulate real-life academic writing tasks usually assigned for academic writing courses. As previously mentioned, the students were instructed to write on an argumentative writing prompt

within a 10-day deadline with access to external resources (e.g., dictionaries, grammar checkers, etc.). The participants took our laptop equipped with the screen recording and completed the task at the location of their choice. Before that, we guided them to run the screen-recording software before their writing activity. At the end of the 10-day deadline, the participants were asked to fetch the laptop where the final text and all the screen-recording data were stored.

Results

Processes

Using the data from our screen recordings and drawing upon Gánem-Gutiérrez and Gilmore's (2018) episode coding categories, we categorized the processes that the writers went through during the two tasks. Based on our modified version of categorization, there were five general episodes the writers underwent while writing:

- *Text construction:* Producing new text in the word processor
- *Re-reading:* Rereading their previously written text (as evident from eye-tracking data)
- *Revising:* Modifying the previously written text
- *Pausing:* Not being involved in any of the above-mentioned activities
- *Use of external resources (only in Non-exam setting):* Using external resources (e.g., online thesaurus, spell/grammar-checkers, etc.)

Using this coding scheme, one of the authors watched all the screen recordings, and the episodes were coded manually; the time spent on each episode was calculated. Finally, the sum of time spent on each episode was tabulated using MS Excel. Tables 2 and 3 show the duration of episodes for each process type in the Exam and Non-exam settings and their proportion regarding the overall time of the task.

Table 2

Duration of Episodes for Each Process Type in Exam Settings (Overall time for all the students was 40 minutes)

	Text construction	Re-reading	Revising	Pausing
Abud	47%	14%	21%	18%
Pari	43%	13%	20%	24%
Ali	35%	15%	19%	27%
Pablo	49%	12%	11%	28%
Wei	51%	9%	11%	29%
Lio	56%	10%	13%	21%
Sophie	48%	13%	23%	16%

Table 3

Proportional Duration of Episodes for Each Process Type in Non-exam Settings

	Text construction	Re-reading	Revising	Pausing	Use of External Resources	Overall time (minutes)
Abud	30%	17%	26%	28%	9%	92 minutes
Pari	33%	15%	20%	21%	11%	135 minutes
Ali	37%	11%	19%	27%	6%	85 minutes
Pablo	34%	16%	23%	23%	4%	73 minutes
Wei	39%	11%	20%	21%	9%	102 minutes
Lio	44%	9%	18%	18%	11%	94 minutes
Sophie	34%	16%	23%	15%	12%	128 minutes

As shown in Tables 2 and 3, Lio spent the highest time on text construction in both tasks while Ali and Abud spent the least in the Exam and Non-exam settings respectively. Also, for re-reading, Wei and Lio spent the lowest time in both tasks, and Ali and Abud spent the highest time in both tasks, respectively. An interesting observation is that Pablo spent the lowest time (4%) on using external resources while Sophie spent 12% of her time on it. This is interesting because they both spent a similar amount of their time on text construction, re-reading, and revising.

Table 4

Frequency and Types of the Queries for the Use of External Resources

Type of the Query	Abud	Pari	Ali	Pablo	Wei	Lio	Sophie
Thesaurus	4	6	2	0	2	3	8
Collocation	5	4	0	0	0	0	5
Translation	0	1	3	3	3	1	0
Grammar check	5	7	0	2	2	2	2
Preposition check	2	3	2	0	1	2	5
Writing Pattern	0	2	0	2	1	1	2
Word usage check	0	2	0	0	0	0	3
Grammar Study	0	1	0	3	1	0	0
Total	16	26	7	10	10	7	25

Another observation is the frequency and type of the queries carried out by the writers during the episode of using external resources. The queries were counted using the screen recordings and were categorized based on a modified version of Yoon's (2016) categorization. As shown in Table 4, Pari and Sophie had the highest, and Ali and Lio had the lowest number of queries compared to the other writers. Also, consulting with thesauruses and grammar checking are, respectively, the most common use of external resources among the writers.

Performances

Coh-Metrix Analysis. Coh-Metrix was used to analyze their underlying features of the collected essays in terms of the mentioned indices. The results from Coh-Metrix are summarized in Table 5.

Table 5

Coh-Metrix Indices for the Essays Written by the Participants in the Exam-setting and the Non-exam Setting

		Abud		Pari		Ali		Pablo		Wei		Lio		Sophie	
		Exam	Non-exam	Exam	Non-exam	Exam	Non-exam	Exam	Non-exam	Exam	Non-exam	Exam	Non-exam	Exam	Non-exam
Descriptive	Word Count	295	312	273	292	241	245	242	238	225	235	231	246	240	259
	Word Length	1.53	1.69	1.81	1.89	1.77	1.81	1.42	1.44	1.49	1.47	1.38	1.51	1.50	1.55
	Sentence Length	24.3	23.8	22.5	24.1	19.7	19.8	20.2	20.2	21.4	21.7	18.8	21.3	21.2	21.8
Lexical Complexity	Word Familiarity*	561	556.3	568.2	561.4	597.1	597	585.3	585.2	579	581.1	595.6	582.9	571.8	568
	Lexical diversity	0.83	.89	0.75	0.79	0.60	0.61	.63	0.65	0.60	0.61	0.58	0.61	0.70	0.74
	Word frequency*	2.09	2.01	2.16	2.01	2.43	2.43	2.37	2.35	2.39	2.41	2.34	2.27	2.30	2.21
	Content Words Concreteness*	346.6	343	333.5	339.6	357.1	356.2	361	360.5	354.8	352.1	368.1	359.2	353.6	349.9
Syntactic Complexity	Embedded Clauses	8.58	8.96	4.07	4.43	3.28	3.28	4.18	4.18	3.98	4.26	3.85	4.19	3.98	4.53
	Number of modifiers per noun phrase	1.04	1.25	0.91	0.96	0.86	0.89	0.83	0.82	0.85	0.87	0.82	0.88	0.91	0.97
	Syntactic Similarity*	0.089	0.082	0.086	0.071	0.085	0.085	0.084	0.086	0.087	0.085	0.083	0.085	0.070	0.081
	Minimal Edit Distance, part of speech	0.69	0.72	0.63	0.67	0.59	0.60	0.60	0.60	0.61	0.59	0.58	0.63	0.61	0.67
Cohesion	Tense and Aspect Repetition	0.72	0.71	0.74	0.76	0.69	0.71	0.68	0.67	0.71	0.70	0.66	0.71	0.68	0.69
	Content Word Overlap	0.034	0.029	0.049	0.042	0.075	0.070	0.057	0.051	0.054	0.049	0.057	0.048	0.047	0.040
	Connective Incidences	118.46	123.12	102.34	109.75	95.24	97.35	98.13	97.35	87.50	91.84	88.53	92.76	91.12	96.04
Accuracy	Error-free T-units divided by total T-units	0.95	1	0.92	0.97	0.78	0.81	0.91	0.91	0.87	0.90	0.81	0.93	0.82	0.89

To compare the indices, due to our limited sample size, non-parametric tests of Wilcoxon-Signed Rank and Chi-Square were run. The results are summarized as follows.

a) Length: For the first three textual indices, separate analyses of chi-square (only for the Word Count) and Wilcoxon-Signed Rank test for Word and Sentence

Lengths were conducted. Based on the results, although the ESL writers used more words in their writings in the Non-exam task (n = 1827, residual = 40, expected = 1787) compared to the exam task (n = 1747, residual = -40, expected = 1787), the results of the chi-square test (χ^2 (1) = 1.79, p > .05, Cramer's V = .022 representing a weak effect size) indicated that the difference was not significant. In terms of word length, the ESL writers had a slightly higher median score on the second task (Mdn = 1.55) than the first task (Mdn = 1.55). The results of Wilcoxon-Signed Ranked test (Z = -2.11, p < .05, r = .179 representing a weak effect size) indicated that there was a significant but weak difference between ESL writers' performance on the first and second tasks in terms of word length. Regarding the sentence length index, while the ESL writers had a slightly higher median score on the second task (Mdn = 21.70) than the first task (Mdn = 21.20), Wilcoxon-Signed Ranked test (Z = -1.57, p > .05, r = .143) showed that there was not any significant difference between ESL writers' performance on the first and second tasks.

b) Lexical Complexity: A non-parametric Wilcoxon-Signed Rank test was run to compare the ESL writer's performance in the first and second tasks in terms of lexical complexity. The ESL writers had a slightly higher median score on the Non-exam task (Mdn = 234.20) compared to the first task (Mdn = 234.06) in terms of lexical complexity. The results (Z = -2.36, p < .05, r = .250 representing a weak effect size) indicated that there was a significant but weak difference between ESL writers' performance on the first and second tasks in terms of lexical complexity.

C) Syntactic Complexity: Based on the results of a non-parametric Wilcoxon-Signed Rank test, the ESL writers had a slightly higher median score on the second task (Mdn = 1.45) than the first task (Mdn = 1.39). The results (Z = -2.20, p < .05, r = .214 representing a weak effect size) indicated that there was a significant but weak difference between ESL writers' performance on the first and second tasks in terms of syntactic complexity.

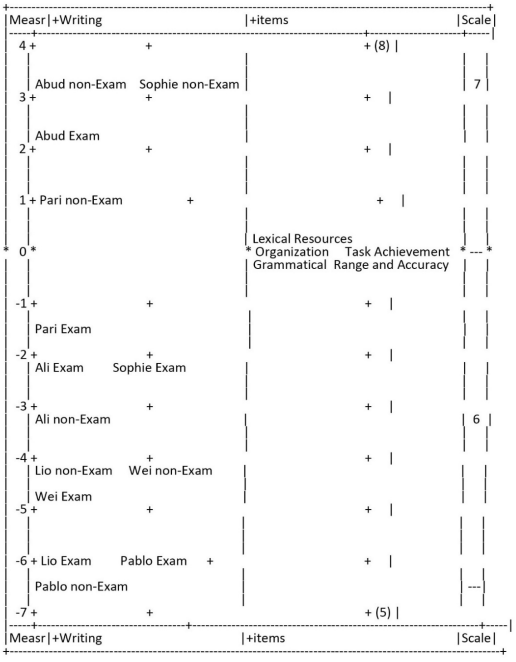
d) Cohesion: A non-parametric Wilcoxon-Signed Rank test showed that the ESL writers had a slightly higher median score on the second task (Mdn = 32.69) than the first task (Mdn = 32.00) in terms of cohesion indices. The results (Z = -2.19, p < .05, r = .179 representing a weak effect size) indicated that there was a significant but weak difference between ESL writers' performance on the first and second tasks in terms of cohesion.

e) Accuracy (error-free T-units): Another non-parametric Wilcoxon-Signed Rank test showed that the ESL writers had a slightly higher median score on the second task (Mdn = .91) than the first task (Mdn = .87) concerning error-free T-units. Based on the results ($Z = -2.21$, $p < .05$, $r = .214$ representing a weak effect size), there was a significant but weak difference between ESL writers' performance on the first and second tasks regarding error-free T-unit length.

Many-Facet Rasch Measurement. To compare the performances of the writers, the raw scores, on a 9-point scale, assigned to the essays by the raters were submitted to FACETS. In our model, we used three facets: raters (n=3), writings (seven students in two tasks, n=14), and the items (four criteria: Lexical Resources, Grammatical Range and Accuracy, Organization, and Task Achievement). However, for a cleaner representation, we do not report the measures related to the raters (after checking the inter-rater reliability). Fig.1 represents the relationship between the two facets of the model.

Fig. 1

The Relationship Between the two Facets of the Model



In this figure, technically referred to as Vertical Rulers, the left column is

the measurement ruler, labeled Measr. The values of this ruler are in logits, ranging from −7 to +4. The column labeled + Writing represents the quality of the writings, ranging from −6 to +3. This means that Abud Non-exam and Sophie Non-exam enjoyed the highest quality, and Pablo Non-exam showed the lowest quality among the writings. The column labeled + Items represents item difficulty based on the four rating criteria. It shows that students performed the best in the aspect of Lexical Resources, and the worst in terms of Grammatical Range and Accuracy, though the differences are very minimal, and their performance in Organization was on par with that of Task Achievement. Finally, the rightmost column shows the IELTS 9-point rating scale, ranging from 0 to 9.

Table 6

Student Writings Measurement Report

Rater	Total score	Logit	Error	Infit	Outfit
				MnSq	MnSq
Abud Non-exam	84	6.77	1.11	2.29	2.88
Sophie Non-exam	84	6.77	1.11	0.08	0.06
Abud Exam	83	5.69	0.95	0.74	0.44
Pari Non-exam	81	4.36	0.73	0.9	0.70
Pari Exam	74	0.70	0.84	0.89	0.48
Ali Exam	73	-0.09	0.95	0.64	0.3
Sophie Exam	73	-0.09	0.95	1.07	1.47
Ali Non-exam	72	-1.09	1.06	0.09	0.07
Wei Non-exam	71	-2.10	0.94	2.07	2.93
Wei Exam	70	-2.85	0.81	1.16	0.93
Lio Non-exam	70	-2.85	0.81	0.75	0.49
Pablo Exam	66	-4.90	0.69	0.76	0.68
Lio Exam	66	-4.90	0.69	1.13	1.23
Pablo Non-exam	65	-5.40	0.71	0.66	0.63

Table 7

Contrast Between Measurement Reports of Exam and Non-exam Writing for Each Student

	Rank in the Exam Task	Rank in the Non-exam Task	Ranking Change	Measure Contrast	t	P value
Abud	1st	1st	0	+0.98	0.70	0.24
Pari	2nd	3rd	-1	+2.64	1.88*	0.036
Ali	3rd	4th	-1	-0.93	0.67	0.25
Sophie	4th	2nd	+2	+4.90	3.52*	0.009
Wei	5th	5th	0	+0.68	0.58	0.28
Pablo	6th	7th	-1	-0.34	.50	0.31
Lio	7th	6th	+1	+1.93	1.65	0.056

** Significant difference at α=0.05(df=22)*

FACETS also produces detailed reports about the performance of individual writers in terms of total scores and logits (Table 6). The writings are ordered from the highest quality, on top, to the lowest quality, at the bottom of the table. Thus, Abud Non-exam and Sophie Non-exam were equally the best essays of all at +6.77 logits and a total score of 84. Moreover, Pablo Non-exam was showed the lowest quality at −5.40 logits and a total score of 65.

Table 7 shows the relative ranking of the students' writings in the two tasks and shows the changes in their standings from Exam to Non-exam. As shown in the table, Abud's writings were ranked first in both tasks and thus have shown no change in the rankings. The table also shows the contrast between the logit measures of the students' writing in both tasks. In Abud's case, the FACETS analysis shows +.98 of logits difference, meaning his Non-exam writing is better at 0.98 logits. To examine the significance of measure contrasts between the two tasks, separate paired t-tests are run using the following formula:

$$t = \frac{E \& NE\ Measure\ Contrast}{\sqrt{SET1^2 + SET2^2}}$$

Where E & N stand for Exam and Non-exam, and SET1 and SET2 stand for Standard Error of Measurement for Exam and Non-exam. Accordingly, the measure contrast between Abud's Exam writing and Non-exam writing is not significant (p = .24). Conversely, although Pari's writing in Non-exam has ranked

lower than her Exam writing, she has performed better in Non-exam at 2.64 logits which is statistically significant (p =.003). Ali's writing in Non-exam is also ranked lower than his writing in Exam, and the measure contrast of -0.93 shows a decrease in the quality of his writing, though the difference is not significant (p =.25). Sophie's writing, ranked 4th in the first task, stood 2nd in the second task with a high measure contrast of 4.90. This sharp increase in the quality of her writing (p = .009) explains Parisa's demotion in Non-exam writing despite its improved quality. Wei also performed slightly better in the second task at .68 logits (p = .28) but gained a similar standing in both tasks. Pablo, on the other hand, was demoted in the second task and had a slightly worse performance in the Non-exam setting (p = .31). Finally, Lio could promote his ranking over Pablo's in the Non-exam writing and showed a better performance compared to Exam writing at 1.93 logits, but the difference falls slightly lower than significant (p = .056).

Discussion

This study aimed to compare the performance of ESL students in the Exam writing situation with their real-life academic writing assignments. Based on the Coh-Metrix indices, the differences between the students' performance in the Exam and Non-exam tasks were significant but weak. This contradicts the findings of Riazi's (2016) study which showed similar performances for students doing two TOEFL writing tasks and one academic writing assignment. They reported similar indices for TOEFL and academic assignments on 15 measures of textual features including syntactic complexity (four measures), lexical sophistication (five measures), and cohesion (six measures). Riazi concluded that "the textual features of the texts produced in the test situation are not significantly different from those produced in the real-life academic writing" (p. 21). However, Riazi does not explicate the details of the process of real-life academic writing in his study. For example, the extent to which the students could or did consult the external resources remains unclear to us.

Based on the results, the writers generally performed better in Non-exam compared to Exam. An average of 0.4 increase from Exam to Non-exam in the IELTS scores given by the raters was observed for the students. This was less than what we expected according to Khuder and Harwood's (2015) 0.8 observed gain

from Exam to Non-exam writing. The results also suggested that the students did not benefit equally from the extra time and resources to improve the quality of their writings. As indicated by our findings in the FACETS analysis, while the overall quality of two students' writing in Non-exam significantly improved, three students' performance was enhanced only marginally, and two students underperformed in a real-life compared to the exam setting. Notably, the relative standing of the quality of the students' writing changed in the two tasks.

The difference in the gained score employing more time and external resources could be explained by the differences in the strategies the writers implemented to achieve their task goals. Given the stakes of exam situations, some students might employ specific strategies to gain a higher score in the exam which are not necessarily constructive for their real-life writing endeavors. Previous research has shown that students who are striving to get a high score on a test like IELTS, benefit from special coaching to improve their scores (Marefat & Heydari, 2018). Likewise, Pennycook (1996) reported that in the context of China the students were encouraged to practice writing on topics expected to appear in the test or memorize texts produced by renowned scholars, and use them in their writing when the topic is relevant. Such studies hint at cases where the applicants who are not really good at academic writing do very well in their tests. Similarly, in an interview conducted by Furneux (2013) one IELTS candidate mentioned that he was good in the IELTS exam but "rubbish" at real academic writing.

The analysis of the time spent by each student on different episodes of their writing as well as the types and frequencies of their queries on external resources bring to mind that the students who were successful in improving the quality of their writing spent more time on using external resources and carried out more queries during this episode. This is in agreement with Roca de Larios, et al. (2008)'s findings who concluded that the writers with different proficiency levels devote varying proportions of time to the writing processes, and the more skilled writers are more likely to regulate their composition processes. Also, similar to Khuder and Harwood's (2015) findings, it can be argued that the distribution of writing processes might have affected the quality of the essays.

A notion that can be taken into account in this regard is the notion of affordances (Hafner & Candlin, 2007; Yoon, 2016). From this perspective, what a

person does, depends on their abilities, goals, values, beliefs, and prior experience (Norman, 2013). An interesting study that draws on the notion of affordances in writing is Hafner and Candlin's (2007) study that specifically monitored the use of a language reference tool to improve L2 writing. The researchers observed the use of a specialized corpus by law students to support their legal writing assignments. The apprentice lawyers, their study implied, used the corpus mostly for legal document searches rather than for lexical or grammar patterns. Although the corpus was provided for them to help them with their word choice and grammar, the tendency of the writer to use it for legal support showed how their identity as lawyers and their membership of that culture strongly influenced their affordance. It should also be considered that due to the difference between real-life and test situations, the writers may tend to have a different conceptualization of the task. Curry (2004), for example, found that in the test situation, writers focus on the word level (grammar and vocabulary choice) rather than idea generation and argumentation as they do not have any resources. However, the writers can be more focused on such aspects of their writing when the material they require is available.

In a longitudinal study of six ESL writers' web-accessed corpus tool use, Yoon (2008) found that the frequency and range of corpus consultation, the types of strategies, and analyses employed by the participants were mostly determined by multiple factors ranging from individuals' prior experiences to disciplinary characteristics. One valuable insight from the study is that learners' motivation to use the corpus technology is determined by the extent to which they have meaningful engagement with it in the process of performing their real-life writing tasks.

The findings of our study corroborate the findings of Yoon (2008) which revealed that the participants' attitudes toward using reference resources as writing assistance were widely different. The cross-case analysis conducted in their study showed that the differences were mainly due to the multifaceted interactions of factors related to the text, writer, and context. This is in line with the implications of Ho Yung and Cai's (2020) study who discussed that writers with high English proficiency do not necessarily perform better in the real-life academic situation and their real-life performance depends on a plethora of factors.

A clear implication of this study would be the long-term impact of

students' past experience and training on their future real-life requirements. Understanding such individual differences and their roots can significantly inform learner training (Kormos, 2012) in the use of online reference tools, which has largely been lacking in the L2 writing pedagogy. In the same vein, in a case study of three Italian L2 writers revising their own compositions while consulting a corpus of Italian texts, Kennedy and Miceli (2010) showed how individual learners' attitudes, goals, and computer literacy affected their ability and willingness to use the unique functions of the corpus. The researchers concluded that the functions of corpus consultation should be explicitly taught.

Another potential implication of this study centers around the validity of conventional timed-impromptu writing tests. The observed gap in processes and performances of students between the exam and real-life academic settings raises serious concerns about the validity of such tests. Twenty-first-century students are now digital natives and are at ease with technology and, therefore, using computers and online resources as sources of assistance has now become a norm for them. As a result, students taking writing tests must perceive the relevance of their test experience to their current and future experiences of writing at university (Chan et al., 2017). In addition, it has been long believed that the restrictions imposed on the writers affect all of them equally and as Worden (2009) mentions, such tests are usually assessed with a 'lower' bar to make it fair for the students. However, as observed in this study, the students might not enjoy equally from lifting the exam restrictions as some of them are already coached for the exam situations and not the real-life venues. Therefore, it is an oversimplification to ignore inherent differences in the psychological characteristics of the writers and the potential differences in the strategies and styles they employ during the complex process of writing. This suggests the need for implementing practical solutions to accommodate the tools like corpora and dictionaries.

The interpretation of the results of this study points to the need for training learners to use technology which should consist of both initial and ongoing scaffolding to implement resources in their L2 writing (Yoon, 2016). In this regard, while the students are taught the shortcuts to improve the quality of their writings in exam settings, they should also be taught to use more general strategies to improve the quality of their writing in real-life tasks.

The findings should, however, be interpreted within the limitations of the study. The first limitation of the study is the number of participants. Due to the highly demanding and time-consuming nature of the tasks involved in this study, despite our attempt to recruit more participants, we were finally able to recruit seven participants. Therefore, the conclusions drawn based on the findings of this study should be considered cautiously. In addition, the weak effect sizes in the Coh-Metrix data analysis and the descriptive nature of some of our observations (e.g., time and frequencies of queries made on using external resources) prevent making strong generalizability claims.

Another complication of the study was to simulate a real-life academic writing setting. Despite the attempts made in the design and execution of the study, we believe asking the writers to record their screen whenever they wanted to write might interfere with the natural process of their writing. Having said that, some students might try to impress the raters while others might not be motivated enough to devote as much time and energy they do in reality. As pointed out by previous studies, when the writers are not responding to an authentic test, they might not be motivated enough (Khuder & Harwood, 2015).

Future research might replicate the current study with bigger sample sizes. In doing so, it is important to compare the students at identical proficiency levels but with different pedagogical backgrounds to eliminate the proficiency variable. Finally, conducting qualitative studies where students are observed in longer intervals and thicker data are obtained from students' writing processes will certainly be revealin

References

Bloom, M. (2008). Second language composition in independent settings: Supporting the writing process with cognitive strategies. In S. Hurd & T. Lewis (Eds.), *Language learning strategies in independent settings* (pp. 103-118). Multilingual Matters.

Chan, S., Bax, S., & Weir, C. (2017). Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS Research Reports Online Series, No. 4*. British Council, Cambridge English Language Assessment and IDP: IELTS Australia.

Conroy, M. A. (2010). Internet tools for language learning: University students taking control of their writing. *Australasian Journal of Educational Technology*, *26*(6), 861-882. https://doi.org/10.14742/ajet.1047

Curry, M. J. (2004). UCLA community college review: Academic literacy for English language learners. *Community College Review, 32*(2), 51-68. https://doi.org/10.1177/009155210403200204

Dziemianko, A. (2012). On the use(fulness) of paper and electronic dictionaries. In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 319-341). Oxford University Press.

Flower, L. S., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365-387. https://www.doi.org/10.2307/356600

Flowerdew, L. (2010). Using corpora for writing instruction. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics (*pp. 444-457). Routledge.

Furneux, C. (2013). What are the academic writing requirements of Masters level study in the Humanities and how far can EAP proficiency tests, such as IELTS, replicate them? Paper presented at the *CRELLA Summer Research Seminar*, United Kingdom.

Gánem-Gutiérrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning*, *68*(2), 469-506.

Hafner, C. A., & Candlin, C. N. (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes, 6*(4), 303-318. https://doi.org/10.1016/j.jeap.2007.09.005

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*(3), 369-388. https://doi.org/10.1177/0741088312451260

Ho Yung, k. W., & Cai, Y. (2020). Do secondary school-leaving English examination results predict university students' academic writing performance? A latent profile

analysis. *Assessment & Evaluation in Higher Education*, *45*(4), 629-642. https://doi.org/10.1080/02602938.2019.1680951

Hyland, K. (2002). *Teaching and researching writing*. Longman.

Kennedy, C., & Miceli, T. (2010). Corpus assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning and Technology*, *14*, 28-44.

Khuder, B., & Harwood, N. (2015). L2 writing in test and non-test situations: Process and product. *Journal of Writing Research, 6*(9)*,* 233-278. https://doi.org/10.17239/jowr-2015.06.03.2

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing*, *21*(4), 390-403. https://doi.org/10.1016/j.jslw.2012.09.003

Lee, S., Lim, G. S., & Basse, R. (2021). The effect of additional time on the quality of argumentation in L2 writing assessment: A mixed-methods study. *Language Assessment Quarterly, 18*(3), 253-272. https://doi.org/10.1080/15434303.2021.1872080

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, *5*(3), 285-337. https://doi.org/10.17239/jowr-2014.05.03.3

Marefat, F., & Heydari, M. (2018). English writing assessment in the context of Iran: The double life of Iranian test-takers. In T. Ruecker & D. Crusan (Eds.), *The politics of English Second Language writing assessment in global contexts* (pp. 67-71). Routledge.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27*(1), 57-86. https://doi.org/10.1177/0741088309351547

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 188-205). IGI Global.

Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books.

Oh, S. (2019). Second language learners' use of writing resources in writing assessment, *Language Assessment Quarterly*, *17*(1), 60-84. https://doi.org/10.1080/15434303.2019.1674854

Pennycook, A. (1996). Borrowing others' words: Text, ownership, memory, and plagiarism. *TESOL Quarterly, 30*(2), 201-230. https://doi.org/10.2307/3588141

Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing*, *28*, 15-27. https://doi.org/10.1016/j.asw.2016.02.001

Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing, 17*(1), 30-47. https://doi.org/10.1016/j.jslw.2007.08.005

Serror, J. (2013). Screen capture technology: A digital window into students' writing processes. *Canadian Journal of Learning and Technology, 39*(3), 1-16. https://doi.org/10.21432/T28G6K

Weir, C. J., O'Sullivan, B., Yan, J., & Bax, S. (2007). Does the computer make a difference? Reaction of participants to a computer-based versus a traditional handwritten form of the IELTS writing component: Effects and impact. *IELTS Research Report*, *7*(6), 1-37.

Worden, D. (2009). Finding process in product pre-writing and revision in timed essay responses. *Assessing Writing, 14*(3), 157-177. https://doi.org/10.1016/j.asw.2009.09.003

Yoon, C. (2016). Individual differences in online reference resource consultation: Case studies of Korean ESL graduate writers. *Journal of Second Language Writing*, *32*, 67-80. https://doi.org/10.1016/j.jslw.2016.04.002

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology, 12*(2), 31-48.

Zheng, B., & Warschauer, M. (2017). Epilogue: Second language writing in the age of computer-mediated communication. *Journal of Second Language Writing*, *36*, 61-67. https://doi.org/10.1016/j.jslw.2017.05.014

Zhi, M., & Huang, B. (2021). Investigating the authenticity of computer- and paper-based ESL writing tests. *Assessing Writing*, *50*, 100548. https://doi.org/10.1016/j.asw.2021.100548

Appendix A

Computer Familiarity Questionnaire

(Adapted from Weir et al., 2007)

Name:_____ (Given name) _____ (Surname)

Student number:_____ Gender: M/F (Please circle)

IELTS result (latest): Overall band:_____ Writing band:_____

There are 15 questions about your experiences and opinions of working with computer. Please **CIRCLE** the number to indicate your answer to each question/statement.

		Always	Very often	Often	Once a while	Never
1)	How often is there a computer available to you to use at each of the following places?					
	a) home	5	4	3	2	1
	b) University	5	4	3	2	1
	c) Public library	5	4	3	2	1
	d) Others (please specify): _____	5	4	3	2	1
2)	How often do you use a computer at each of the following places?					
	a) home	5	4	3	2	1
	b) University	5	4	3	2	1
	c) public library	5	4	3	2	1
	d) others (please specify): _____	5	4	3	2	1
3)	How often do you use computer for					
	a) surfing the Internet ?	5	4	3	2	1
	b) electronic communication, e.g. emails?	5	4	3	2	1
	c) study-related activities?	5	4	3	2	1
	d) Others (please specify): _____	5	4	3	2	1
4)	How often do you use the computer **software** for					
	a) games?	5	4	3	2	1
	b) word processing?	5	4	3	2	1
	c) spreadsheets?	5	4	3	2	1
	d) painting or graphics?	5	4	3	2	1
	e) data or text analysis?	5	4	3	2	1
	e) Others (please specify): _____?	5	4	3	2	1

4)	How often do you use the computer **software** for					
	a) games?	5	4	3	2	1
	b) word processing?	5	4	3	2	1
	c) spreadsheets?	5	4	3	2	1
	d) painting or graphics?	5	4	3	2	1
	e) data or text analysis?	5	4	3	2	1
	e) Others (please specify): _____?	5	4	3	2	1
5)	How often do you take a test on					
	a) paper?	5	4	3	2	1
	b) computer?	5	4	3	2	1
		Very comfortable	Quite Comfortable	Comfortable	Quite uncomfortable	Very uncomfortable
6)	How comfortable are you with using a computer in general?	5	4	3	2	1
7)	How comfortable are you with using a computer to write a paper?	5	4	3	2	1
8)	How comfortable are you with taking a test on					
	a) computer?	5	4	3	2	1
	b) paper?	5	4	3	2	1
9)	How do you feel about using the keyboard (typing)?	5	4	3	2	1
10)	It is very important to me to work with a computer.	5	4	3	2	1
11)	To play or work with a computer is really fun.	5	4	3	2	1
12)	I use a computer because I am very interested in this.	5	4	3	2	1
13)	I forget the time, when I am working with the computer.	5	4	3	2	1
		Excellent	Good	Fair	Poor	Very poor
14)	If you compare yourself with other students, how would you rate your ability to use a computer?	5	4	3	2	1

Appendix B

Writing Task Prompts

Writing Task 1 prompt

Environmental pollution has become so serious that many countries are trying to solve them. What are the most serious problems associated with it and what solutions can you suggest?

Writing Task 2 prompt

The internet has transformed the way information is shared and consumed, but it has also created problems that did not exist before. What are the most serious problems associated with the internet and what solutions can you suggest?