An Online Big-Data Driven Design of Reading and Writing Test

Yuwei Sun[1], Yongcheng Wen[2(✉)], and Yazhen Zhu[3]

[1] Columbia University, New York, NY 10027, USA
[2] Shenzhen MSU-BIT University, Shenzhen, Guangdong 518172, People's Republic of China
1120200244@smbu.edu.cn
[3] Royal College of Art, London SW7 2EU, UK

Abstract. This paper presents an online big-data driven design of reading and writing tests, incorporating empirical data analysis. The study aims to investigate the nature of reading and writing abilities, their corresponding relationship, and the impact of background variables on learning-oriented test performance. The counterpart was administered through an online platform, where data are collected for assessing the performance of students. The objectives of our work are to provide insights into the test design, delivery, and feedback mechanisms, and to conduct a statistical evaluation of the test's reliability, validity, and correlations. The findings contribute to our understanding of reading and writing assessment in an online context, while also highlighting the implications of background variables on test performance.

Keywords: Learning-oriented test · Online big-data driven task · Reliability & Validity of the test

1 Introduction

1.1 Motivation

Language proficiency assessment plays a crucial role in language learning and teaching. However, traditional assessment methods often focus solely on measuring students' outcomes without considering their learning progress and individual growth. In contrast, a learning-oriented test model offers a valuable approach that emphasizes the learning process itself and provides formative feedback to support students' improvement.

The objective of this work is two-fold:

- To provide information about the learning-oriented test in terms of the process of design, its test constructs, delivery, and feedback.
- To conduct a statistical evaluation of the reliability and validity of the test and examine whether the test supports its claims.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 335–353, 2024.
https://doi.org/10.1007/978-3-031-65126-7_29

The motivation behind this paper lies in the development and implementation of a learning-oriented achievement test for ESL students at the Upper Intermediate Level. The test aims to assess reading comprehension and writing skills in a professional context, focusing on the theme of "The Art of Complaining" from Unit 10 of the coursebook. In order to ensure the relevance and effectiveness of the test, it is essential to explore the nature of reading and writing abilities on this test and investigate the relationship between these two constructs.

The literature review conducted for this study revealed valuable insights into the relationship between reading and writing abilities in language learning. Studies by Bazerman (1980) and Cumming et al. (2004) highlighted the interconnectedness of reading1 and writing, showcasing how these skills mutually reinforce each other [4, 12]. Additionally, the works of Berninger et al. (2002) and Grabe (1991) contributed to a deeper understanding of how language by hand (writing) and language by eye (reading) are closely intertwined in the learning process [5, 16].

By drawing on the insights from these literature reviews, the test design was strategically developed to create an integrated learning-oriented assessment. The counterpart not only provides students with summative scores but also offers formative feedback, aiding their improvement in both reading comprehension and writing abilities. The design of test aligns with the established theoretical frameworks and aims to support the development of real-world language and communicative competencies.

Through the utilization of an online platform, the test was administered using a big-data driven approach, following the footsteps of Gebril and Plakans (2009) and Plakans & Gebril (2012) [15, 26]. This approach harnessed the benefits of data analysis from student responses, providing valuable insights into student performance and test effectiveness. Furthermore, the statistical evaluation conducted in this study adheres to the principles of validity and reliability highlighted by Borsboom et al. (2004) and Myford & Wolfe (2003) [7, 24].

This paper aims to achieve three primary objectives:

1. To provide a comprehensive overview of the learning-oriented test, including its design process, the test constructs it measures, the delivery methods employed, and the feedback mechanisms in place.
2. To conduct a statistical evaluation of the reliability and validity of the test, meticulously assessing whether the test aligns with its intended claims.
3. To collect the data, we utilized an online platform, adopting a truly online big-data driven approach. This allowed us to benefit from an extensive array of student responses and conduct thorough data analysis, providing valuable insights into student performance and the effectiveness of the test.

The paper will provide detailed information about the test design and present a comprehensive statistical analysis of the results. The discussion section will present our main findings, address any significant limitations, and offer recommendations for future research and test design enhancements.

The following research questions will be addressed in this paper:

1. What is the nature of reading and writing ability on this test?
2. What is the nature of the relationship between reading and writing ability on this test?

3. To what extent were the raters consistent when rating writing ability in this test?
4. What is the nature of the relationship between the test-takers' self-reported length of studying English, and their performance on the test?

1.2 Organization

The organization of this work is as follows. Section 2 will first present procedure of the theoretical conceptualizations for the constructs of the test is measuring, including online big-data collection and feedback system used for a learning-oriented approach, assessing reading abilities and providing detailed feedback to improve writing. In Section 3 , this paper will elaborate on the statistical analysis of the results of the assessment. Section 4 of this work will conclude with a discussion of the main findings, the major limitations, as well as recommendations for future tests or research directions.

2 Online Big-Data Collection and Feedback System

2.1 Online Big-Data Collection: Platform and Procedures

With a learning-oriented approach in mind, twelve multiple-choice questions for the reading passage were designed in a way that they could also serve as the input for the writing assignment. The reading passage was about how to write a polite email to the professor and the test-takers will be assessed in terms of their endophoric literal and endophoric implied reading abilities. The reading passage with salient discourse markers also acts as a socio-transactional and linguistic resource for test-takers for the writing section.

To complete the writing assignment, students had to go through three stages: read, reflect, and write. To begin, students were given a reading passage on how to politely write an email to professors where they need to identify and highlight essential features of the input to finish the reading questions. During the reflection phase, they received feedback for each question once they completed all the multiple-choice items. The first two steps of the process were meant to teach students how to construct a professional and polite email through reading and reflection, as well as to provide them with background knowledge and resources to write their own emails in the second part of the test (Table 1).

A writing scoring rubric based on four components was used to evaluate the students for their writing section. The components took into consideration the language control, the content's accuracy and elaboration, and the rhetorical control. The sociolinguistic appropriateness like tone and formality were incorporated under language control. The scoring results were provided to the students together with feedback on various parts they should improve on. For the actual scoring part, both raters first independently rated the four constructs for each response, and then they averaged their scores on each construct and the total added-up score as well.

Each test-taker received a score report that has both summative scores (i.e., each construct's score and an overall score) and formative feedback (i.e., how their response was scored, a complete version of the rubrics for their reference, interpretations of each construct's score, global comment, as well as suggestions on how their writing could be improved) (Tables 2 and 3).

Table 1. The Test Structure for Unit 10

Test Component	Task type	Number of Items/ Tasks	Time	Scoring
Reading: -Endophric literal (e.g., summarizing the gist, identifying details) -Endophoric implied (e.g., infering)	Selected response: -MC items	12 items	60 min in total	- Dichotomous (0/1) - 12 points in total
Writing: -Language Control -Content Accuracy -Rhetorical Control -Content Elaboration Theme: Polite email in the academic domain	Constructed-response task/ extended production: -Integrated reading-for-writing task with scaffolding			Analytic Scoring with a rubric
				-Rating scale ranging from 1–5 criteria
				- 5 points for each
				- 20 points in total
				- 2 raters

Table 2. The Writing Scoring Rubrics used on this test

Score (Level of Control)	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
5	The language of email is clear, cogent and smooth, and displays high accuracy in grammatical and lexical choices from a wide range. Consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some very minor lexical or grammatical errors which do not interfere with understanding	The email information is accurate and relevant to the task. The form of the email is complete and shows an accurate understanding of the reading input	The email includes clear and logical explanations of the student's situation. The relationships between ideas are supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are persuasive	The email is fully developed with considerations and examples from the reading input. The email may also contain considerable extra background knowledge of filing a complaint or writing a polite email in a professional setting

(*continued*)

Table 2. (*continued*)

Score (Level of Control)	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
4	The language of email is generally clear, cogent and smooth, and displays general accuracy in grammatical and lexical choices from a wide range. Fairly consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some minor lexical or grammatical errors, which may or may not interfere with understanding	The email information is generally accurate and relevant to the task and shows a generally good understanding of the reading input. The form of the email is fairly complete, maybe with minor missing parts	The email includes fairly clear and logical explanations of the student's situation. The relationships between ideas are fairly supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are fairly persuasive	The email is fairly developed with considerations and examples from the reading input. The email might also contain some extra background knowledge of filing a complaint or writing a polite email in a professional setting
3	The language of email is moderately clear and cogent, and displays moderate accuracy in grammatical and lexical choices from a moderately wide range. Somewhat consistent pragmatic appropriateness and politeness in tone, register, and stance, but may contain obvious inconsistency. The email has a noticeable amount of grammatical or lexical errors, which slightly interfere with understanding	The email information is moderately accurate and relevant to the task, and shows some understanding of the reading input. The form of the email is moderately complete, but a few parts are missing	The email includes some clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are somewhat persuasive	The email is somewhat developed with details or considerations from the reading input. The email might also contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting

(*continued*)

Table 2. (*continued*)

Score (Level of Control)	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
2	The language of email is clear or cogent at times but exhibits problems in being consistent. The email displays an inappropriate tone, register, or stance. The language either shows limits in control of vocabulary and grammar, or has a large amount of errors, which significantly interfere with understanding	The email information is somewhat relevant to the task but leaves most of it unattended. The response fails to show understanding or consideration of the reading input. The form of the email is incomplete	The email barely includes any clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are barely persuasive	The email lacks important considerations and examples from the reading input. The email might sometimes contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting
1	The language of email is very limited in coherence or clarity without any pragmatic appropriateness considerations. Language only has discreet words or phrases that barely connect, or is full of lexical and grammatical errors, which severely interfere with understanding	The email information is not relevant to the task. The response fails to show understanding or consideration of the reading input. At this level, the form of the email is either missing important parts or only has few parts	The email has poor and illogical explanations of the student's situation without any logical connectors or/and cohesive devices. The explanations and requests made in the email are not persuasive	The email lacks considerations and examples from the reading input. The email fails to contain extra background knowledge of filing a complaint or writing a polite email in a professional setting

Google forms, an online survey platform, was used to administer the test, allowing users to create customized and complex surveys and receive statistical reports after the data had been collected. In the preceding lecture, students were informed of the test structure and the format.

Eight students in community language program (CLP) Upper-intermediate 3 took the exam. They were given the link to the Google forms test in the Zoom chat box after the test instructions. While taking a test, students were required to remain in the Zoom conference and keep the camera on. Students were also told that the test time was 60 min for both sections of the test, during which they could go back to different sections to check their responses. All students were told to submit their tests when the time was up, and the class then moved on to regular class activities.

Table 3. Item Coding and Keys for the Reading Section.

Observed Variable	Item Number	Answer Key
Endophoric Literal	1	**D**
	3	**C**
	4	**B**
	6	**C**
	7	**B**
	10	**D**
Endophoric Implied	2	**A**
	5	**D**
	8	**B**
	9	**D**
	11	**C**
	12	**A**

The test was designed into 2 parts, reading comprehension, and extended polite email writing. Once the test-takers completed the first part, they received detailed feedback (explanation for different options) for each multiple-choice question that was delivered in Google Forms. After the test-takers finished reading the feedback, they proceeded to the second part to write a formal email to the professor based on what they had learned from part 1. Test-takers can write their response in either the given Google forms or in a Google doc.

2.2 Feedback System

Multiple types of feedback were given to the test-takers in the test, in both summative and formative ways. The first feedback was given in between the two parts, where students could utilize the feedback for their reading section performance to pave the way for the email writing in the second part. After the two raters finished scoring all the responses, they gave each test-taker a score report. The score report included the overall score of the student, their score for each component, explanations on how each construct was rated, edits of their email, together with individualized detailed feedback on how they could possibly improve. The reports were also sent to the class teachers to help them better understand where the students were and how to close their learning gap.

Based on the scores, most of the test-takers showed good comprehension of the information from the reading passage, and most of them successfully composed an email that showed at least moderate understanding or consideration of the reading passage. This test also critically and effectively evaluated how much CLP students comprehensibly learned the grammar focus and the socio-pragmatic knowledge focus from Unit 10 on how to properly initiate a complaint.

Below is an example score report (Table 4).

Table 4. A sample score report for the test

Student M
Score: 22/24 for reading, 19/20 for writing
The email was well-written with a careful selection of politeness hedges (e.g., would you mind, would you please, I would like to…, etc.). The email exhibits a clear and appropriate explanation of the situation and suggested reasonable alternative solutions. The email also shows a good amount of consideration of the reading input. There are some minor grammatical and lexical errors, but they did not interfere with understanding. Please see your Google Doc. for detailed edits.
Please see below for the scoring rubrics.

Score	Language Control	Content Accuracy	Rhetorical Control	Content Elaboration
	Your score: 4	**Your score: 5**	**Your score: 5**	**Your score: 5**
5	The language of email is clear, cogent and smooth, and displays high accuracy in grammatical and lexical choices from a wide range. Consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some very minor lexical or grammatical errors which do not interfere with understanding.	The email information is accurate and relevant to the task. The form of the email is complete and shows an accurate understanding of the reading input.	The email includes clear and logical explanations of the student's situation. The relationships between ideas are supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are persuasive.	The email is fully developed with considerations and examples from the reading input. The email may also contain considerable extra background knowledge of filing a complaint or writing a polite email in a professional setting.
4	The language of email is generally clear, cogent and smooth, and displays general accuracy in grammatical and lexical choices from a wide range. Fairly consistent pragmatic appropriateness and politeness in tone, register, and stance. It may include some minor lexical or grammatical errors, which may or may not interfere with understanding.	The email information is generally accurate and relevant to the task and shows a generally good understanding of the reading input. The form of the email is fairly complete, maybe with minor missing parts.	The email includes fairly clear and logical explanations of the student's situation. The relationships between ideas are fairly supported by appropriate logical connectors and cohesive devices. The explanations and requests made in the email are fairly persuasive.	The email is fairly developed with considerations and examples from the reading input. The email might also contain some extra background knowledge of filing a complaint or writing a polite email in a professional setting.
3	The language of email is moderately clear and cogent, and displays moderate accuracy in grammatical and lexical choices from a moderately wide range. Somewhat consistent pragmatic appropriateness and politeness in tone, register, and stance, but may contain obvious inconsistency. The email has a noticeable amount of grammatical or lexical errors, which slightly interfere with understanding.	The email information is moderately accurate and relevant to the task, and shows some understanding of the reading input. The form of the email is moderately complete, but a few parts are missing.	The email includes some clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are somewhat persuasive.	The email is somewhat developed with details or considerations from the reading input. The email might also contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting.
2	The language of email is clear or cogent at times but exhibits problems in being consistent. The email displays an inappropriate tone, register, or stance. The language either shows limits in control of vocabulary and grammar, or has a large amount of errors, which significantly interfere with understanding.	The email information is somewhat relevant to the task but leaves most of it unattended. The response fails to show understanding or consideration of the reading input. The form of the email is incomplete.	The email barely includes any clear and logical explanations of the student's situation. The email exhibits almost no logical connectors or cohesive devices. The explanations and requests made in the email are barely persuasive.	The email lacks important considerations and examples from the reading input. The email might sometimes contain a limited amount of extra background knowledge of filing a complaint or writing a polite email in a professional setting.
1	The language of email is very limited in coherence or clarity without any pragmatic appropriateness considerations. Language only has discreet words or phrases that barely connect, or is full of lexical and grammatical errors, which severely interfere with understanding.	The email information is not relevant to the task. The response fails to show understanding or consideration of the reading input. At this level, the form of the email is either missing important parts or only has few parts.	The email has poor and illogical explanations of the student's situation without any logical connectors or/and cohesive devices. The explanations and requests made in the email are not persuasive.	The email lacks considerations and examples from the reading input. The email fails to contain extra background knowledge of filing a complaint or writing a polite email in a professional setting.

3 Data Analysis

3.1 Results for the Reading Task

As the results in Table 5 showed, the reading section had 12 multiple choice questions, for a full score of 12. The mean was 10.00, the median was 10.00, the mode was 9, and the standard deviation was 1.069. The skewness was .935. The standard error of skewness was .752, and the kurtosis was .350, with a standard error of kurtosis of 1.481.

Table 5. Descriptive Statistics for the Reading

	Central Tendency						Dispersion		Frequency	Distribution
Read Tot	N	Mean	Mode	Median	Min	Max	Range	SD	Skewness	Kurtosis
	8	10	9	10	9	12	3	1.069	.935	.35

The measures of distribution, kurtosis and skewness were included to measure comparability to a normal distribution, based on the assumption that a large enough group was being tested. However, it is important to remember that our sample size (N = 8) is very small so the distribution and characteristics were only situated within this study's context. The positive skewness of .935 for the reading total indicates that there were more lower scores than higher scores. If we only look at the skewness, we might conclude that having more lower scores is not ideal since the test was designed to be an achievement test, so we wanted more people to have a high score than otherwise.

However, a mean and median score of 10 out of 12 indicates that most of the test-takers did a fairly good job. It suggests although some test-takers performed better, the test-taker did a good job of reading overall. The standard deviation of 1.069 and kurtosis of 1.481 were both within the normal range, suggesting a moderate to fairly narrow distribution of scores. This means that most of the reading scores of the test-takers were similar. The internal consistency reliability and standard error of measurement (*SEM*) were calculated for all variables. Taking the composite variables for each component of the data, the internal consistency reliability was measured by Cronbach's alpha. Alpha is calculated by using the individual item-level data for all the items (Table 6).

Table 6. Reliability of the reading (N = 8)

Cronbach's alpha (α)	Number of MC Items
− 0.170	12

Note. The alpha(α) is negative due to a negative average covariance among items. This violates reliability model assumptions. The reliability will thus be considered .00

Cronbach's alpha for our test is −.170, indicating a negative average covariance among items and violating reliability assumptions. Ideally, alpha should be .70 or higher

for low-stakes classroom assessments. Due to a small sample size, normal distribution of scores might be compromised. As alpha cannot be below 0, we consider it as .00. Another measure of internal reliability, the standard error of measurement (*SEM*), accounts for sample variability and measurement error. It is calculated using the square root of "1" minus Cronbach's alpha, multiplied by the standard deviation (*S*) for the task.

$$SEM = \mathbf{SEM} = S\sqrt{1 - r'_{xx}} \tag{1}$$

$$S = \mathbf{standard\ deviation}(S = 1.069) \tag{2}$$

$$r'_{xx} = \mathbf{reliability}(r'_{xx} = 0.00) \tag{3}$$

The SEM value based on the standard deviation (*S*) of 1.069 is also 1.069. Using a 68% confidence interval, the passing cut-score for the reading part is 8.4 out of 12. Scores above 9.469 ($1SEM$) can be 68% confident of passing, while scores below 7.331 ($-1SEM$) can be 68% confident of not passing. Due to the low alpha, we cannot confidently determine pass or fail for scores between $-1SEM$ and $+1SEM$ (Carr, 2011). For the internal consistency analysis of the 12 multiple-choice reading questions, we examined Item Facility (IF), Discrimination Index (DI), and Alpha if Deleted. The ideal IF range is .3-.7. Positively discriminating items (DI $\geq$ .4) are usually desirable, while negatively discriminating items should be revised or deleted (Carr, 2011) (Table 7).

Table 7. Item analysis of the Multiple-Choice Questions

Item	Item Facility	Discrimination	Cronbach's alpha if removed	Decision
Endoliteral 1	.63	−.228	.079	Revise
Endoimplied 2	.38	.038	−.319[a]	Keep
Endoliteral 3, 7	1	0	−.319[a]	Keep
Endoimplied 4, 5, 6				
Endoimplied 8	.88	−.314	.07	Revise
Endoimplied 9	1	0	.07	Keep
Endoliterate 10	.63	.038	−.319[a]	Keep
Endoimplied 11	.88	.051	−.273[a]	Keep
Endoimplied 12	.63	.038	−.319[a]	Keep

Note. ([a])The value is negative due to a negative average covariance among items. This violates reliability model

The items in the test show a wide range of difficulty, with Item Facility (IF) ranging from .38 to 1.00. Six items were considered "no variance" as they were answered correctly by all test-takers (IF = 1), indicating they might be too easy. Despite this, we decided to keep all non-discriminating items due to the low-stakes nature of the test in

a classroom-based language teaching context. D-values range from −.314 to .051, suggesting limited discrimination ability, likely influenced by the small sample size (N = 9) and narrow range of total reading scores (9–11). Two items (endoliteral 1 and endoimplied 2) with negative D-values will be revised, while four items with low D-values will be kept, considering the small sample size and low-stakes nature of the test. The negative calculated alpha indicates issues with the reliability model assumptions, so decisions are not solely based on alpha. Distractor Analysis was conducted to examine why some items failed to differentiate between high and low-performing test-takers. Representative items (endoliteral 1 and endoimplied 9) were selected for further investigation. Endoliteral 1, with a negative D-value, will be removed to improve Cronbach's Alpha, and endoimplied 9, with an IF of 1 and D-value of 0, was found to be non-discriminating (Table 8).

Table 8. Distractor Analysis for endoliteral 1 (total test-takers N = 6, high and low-performing group K = 6)

Answer Choice	Frequency	High (n = 3)	Low (n = 3)	Item Facility (Upper)	Item Facility (Lower)	Item Discrimination
A, B	0	0	0	0	0	0
C	2	1	0	.333	0	.111
D*	4	2	3	.667	1	−.111

*. *Key*

The IF of .63 shows 63% answered correctly, but the Dd-value of −.228 reveals an issue, favoring low-performers. Deleting the item increases Cronbach's Alpha to .079 (−.170 originally due to small sample size). Table 13 indicates negative discrimination for Choice D (key) with only two high-performers selecting it, while all three low-performers did. Choices A and B were poor distractors with no selections. In conclusion, all choices need revision. A and B should be more distracting, and C and D must better discriminate between high and low performers, measuring endophoric implied understanding accurately.

Table 9. Distractor Analysis for endoimplied 9 (total test taker N = 6, high and low-performing group K = 6)

Answer Choice	Frequency	High (n = 3)	Low (n = 3)	Item Facility (Upper)	Item Facility (Lower)	Item Discrimination
A, B, C	0	0	0	0	0	0
D*	6	3	3	1	1	0

*. *Key*

As shown in Table 9, all the test-takers (all the high- and low-performing group and the people in the middle) chose the right answer D, making the Item Facility 1.00, which indicates this question might be too easy. But considering the nature of the test is a low-stakes classroom-based language learning achievement test, we decided to keep the item. The D-value of the item is 0, which means it cannot discriminate between low performing test-takers from high performing ones. No test-takers chose any of the other three distractors, indicating that the distractors may need to be revised had we decided to revise this item. We propose to revise the items in the following way had we decided to: A, and B, and C need to be more distracting so they can properly discriminate between the high and low performing groups. Again, as previously stated, the purpose of revision is to make sure the item is actually assessing what it claims it measures, which is the endophoric implied reading ability of the student (Table 10).

Table 10. Stem and Leaf Plot of results for Reading task

Score	9.00	10.00	11.00	12.00
Frequency	XXX	XXX	X	X

Stem Width: 1.00.
Each Leaf: 1 case(s).

The reading part of the test consists of two variables: endophoric literal and endophoric implied. The 12 multiple-choice items were evenly split, with six for each variable. We used the Pearson Product-Moment formula to analyze the scores and assess the relationship between the two item types. Positive correlation between them is expected, as they measure the same reading ability. A correlation coefficient above .75 is high, .5–.74 is moderate, .25–.49 is low, and 0 means no correlation.

Table 11. Pearson Correlation Matrix for reading total: Observed Variables (N = 8)

Endo Lit Total		
Endo Implied Total	Person Correlation	−.114
	Sig. (2-tailed)	.788

The result from Table 11 showed a slightly negative correlation between total endophoric literal and total endophoric implied scores for each student. This means that the better a test taker did in the endophoric literal questions, the poorer he or she did in the endophoric implied questions, and vice versa. But this correlation was not found to be statistically significant, which means this could be a chance phenomenon. Based on the data, we may conclude that there is not sufficient evidence to suggest the two types of questions were measuring the same ability as they were supposed to. A few reasons could explain this. One reason could be that the reading test per se had low reliability, and the items need to be improved to serve as a sufficient condition to

support the claimed validity of the test. Also, the sample size of the study (N = 8) was extremely small, and the total reading score range of students was very small (9 was the lowest and 12 was the highest), meaning the data was negatively skewed. On the bright side, if all the reading scores of test-takers were negatively skewed, it means most of them did very well in the test. This could be likely attributed to the fact that some of the questions reflected the content taught in class. Good and bad email examples of writing were analyzed and discussed in the class, so maybe by the time the test was administered, the students already had a good understanding of the dos and don'ts of writing emails to professors.

3.2 Results for the Writing Task

Table 12 showed writing performance of test-takers in language control, content accuracy, content elaboration, and rhetorical control, four of which were scored from 1 to 5, respectively. Within a total possible score of 5, the mean was 4.36, the median was 4.5), and the mode was 4.875. The standard deviation was .504. The skewness was −.456, with a standard error of skewness of .752. We can also read that the kurtosis was −1.714 and the standard error of kurtosis was 1.481. Since the minimum for the writing was 3.635 and the maximum was 4.875, the range was 1.24.

Table 12. Descriptive Statistics for the Writing

	Central Tendency						Dispersion		Frequency	Distribution
Write Ave	N	Mean	Mode	Median	Min	Max	Range	SD	Skewness	Kurtosis
	8	4.36	4.875	4.5	3.635	4.875	1.24	0.504	−.456	−1.714

The negative skewness of −.456 indicates that higher scores occurred fairly more frequently than lower scores. A negative skewness is desirable in this case since the test was designed to be an achievement test so we wanted to see more high scores. It suggests that most of the test-takers did a fairly good job in successfully completing the email writing task. Since the two parts of the test were designed in a learning-oriented way where test-takers needed to incorporate the reading input into writing tasks, this negative skewness could also be interpreted that a big proportion of the test-takers well comprehended the reading materials and successfully incorporated the understanding into their writing. The mean score of 4.361 and the median score of 4.50 (out of 5) also suggested that the test-takers did a good job in writing overall. The standard deviation of 0.504 and kurtosis of −1.714 suggest a standard distribution of scores. To put it in a different way, scores of test-takers showed some moderate variation in writing (refer to Table 13 and 14). (Fig. 1)

Internal Consistency Reliability for the Writing Task was calculated using Cronbach's Alpha and was done in a similar way when Internal Consistency was calculated for the Reading MC items. The difference is unlike the Reading MC scores which were dichotomous, Writing Task's Internal Consistency was estimated based on the average composite variables (between rater 1 and 2) for each of the four components (Language

Table 13. Stem and Leaf Plot of results for Writing task

Score	3.75	4.00	4.25	4.50	4.75	5.00
Frequency	XX	X	X		XX	XX

Stem Width: .25.
Each Leaf: 1 case(s).

Table 14. Reliability of the writing task (N = 8)

Cronbach's Alpha (α)	Number of components of rubric
.888	4

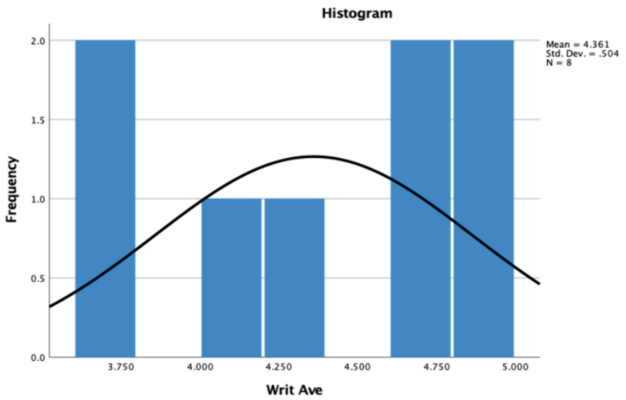

Fig. 1. Histogram of the results of the writing task with a normal distribution

Control, Content Accuracy Control, Rhetoric Control, and Content Elaboration Control), and the calculations were conducted using interval variables. Table 15 shows the alpha was estimated to be .888.

A .888 reliability can be interpreted as that only 11.2% of the observed variance can be attributed to measurement errors or other unaccounted factors but 88.8% of the score variance is attributed to true score variance. Internal consistency reliability for high-stakes standardized tests like TOEFL usually needs to be extremely high. TOEFL claims to have an overall reliability estimate of 0.95 and a 4.26 SEM (standard errors of measurements) (ETS, 2018), which can be seen as high consistency and reliability since the smaller the value of SEM, the higher the measurement quality, and the more precise the test scores would be. For low-stakes classroom-based assessments like our test, .70 is considered an acceptable threshold (Carr, 2011), and thus we can consider .888 to be a very high reliability in this case. In CTT, an alpha of .888 is interpreted as 88.8% of variance attributable to true test-taker ability, and 11.2% of variance attributable to error (Carr, 2011, pp. 108–109). It means the writing section can reliably assess writing

responses of students. Two big reasons that can potentially explain the high reliability are a). The two raters both have spent a considerable amount of time studying, living, or working in the United States with frequent exposure to English as a second language so they might have a very similar overall judgment of the writing of a student; *and* b). The two raters had numerous discussions and rounds of revisions of the scoring rubrics and descriptors to make sure they had reached an agreement on how they would assess the writing response of a test-taker.

The standard error of measurement (*SEM*) was also calculated in this section. The formula is 1 minus r_{xx} (Reliability Estimate) and then calculating the square root of the result. Then the square root is multiplied by S (the Standard Deviation). Below is the formula.

$$SEM = \mathbf{SEM} = S\sqrt{1 - r'_{xx}} \tag{4}$$

$$S = \mathbf{standard\ deviation}(S = .504) \tag{5}$$

$$r'_{xx} = \mathbf{reliability}(r'_{xx} = .888) \tag{6}$$

Based on the descriptive statistics in previous sections, we can know that the standard deviation is .504 and the calculated SEM is 0.169. Considering this is a low-stakes test, we use a 68% confidence interval ($\pm 1SEM$), which means if one of the test takers were to take the test again, we could say with 68% confidence that his or her score is most likely to fall within $\pm 1SEM$ of their current score on the writing task. The cut-score for passing is 70%, and it means we can be 68% confident that test-takers who received below 3.331 did not receive the passing grade for the task. $+1SEM$ is 3.669 or more and this means that we can be 68% confident that those who scored above 3.669 received a passing grade for the task (i.e., 3.5 out of 5 = 70%). For those test-takers who scored between -1 SEM (3.331) and $+1SEM$ (3.669), we cannot say with 68% confidence that they either passed or didn't pass. A 68% confidence level ($\pm 1SEM$; ± 0.169) with 70% as the cut-score would mean a student needs to get at least 3.33 to pass the exam (Tables 15, 16 and 17).

Table 15. Writing scores for all the test-takers

ID	1	2	3	4	5	6	7	8
SCORE	4.375	3.635	4.625	4.000	4.875	3.75	4.75	4.875

Based on the 68% confidence interval, eight test-takers (IDs 1, 2, 3, 4, 5, 6, 7, and 8) received passing grades with scores above 3.669 out of 5. However, one test-taker (ID2) received a score of 3.635, falling within the interval (3.331 to 3.669), making it uncertain to confidently conclude a passing grade within the 68% confidence level. To assess inter-rater reliability, correlations were calculated between two raters' scores for four components (language control, content accuracy, rhetoric control, and content elaboration) using Spearman rank-order and Pearson product-moment.

Table 16. Inter-Rater Reliability Correlation Matrix: Observed Variables (N = 8)

Language control R1, R2	.577
Content accuracy R1, R2	.600
Rhetorical control R1, R2	.667
Content elaboration R1, R2	.775*

*. *p < .05, 2-tailed*

Table 17. Inter-Rater Reliability Correlation Matrix: Writing Average Scores (N = 8)

	Writing Average Rater 2
Writing Average Rater 1	.883**

**. *p < .01, 2-tailed*

The correlation coefficients between the four individual variables ranged from moderately positive (e.g., 0.577) to highly positive (e.g., 0.775). The inter-rater reliability was .577 for language control, .600 for content accuracy, and .667 for rhetoric control, none of which, according to SPSS, was found to be statistically significant at the .05 level, meaning that there is not a significant linear correlation between rater 1 and rater 2 in the sample. The correlation coefficient for content elaboration, however, was .775 and was found to be statistically significant at the .05 level, meaning that there is a 95% chance the correlation is not a chance phenomenon. Turning to the measurements of Inter-rater reliability computed using the composite averages by rater, Table 18 below provides the Pearson correlation yielded from this analysis.

The correlation for the Writing Average Rater 1 and Writing Average Rater 2 was .883 at the .01 level, meaning there is a 99% chance that the correlation is not a chance phenomenon. Considering that the correlation coefficient range is −1 to 1, we consider the $r = .833$ to be a high correlation coefficient, meaning that the two raters were highly congruent in their overall rating for writing performance of students.

The high inter-rater reliability on writing average score of the R1 and R2 might be because the two raters worked together, or 'norming', through multiple rounds of discussions and revisions on the test design, scoring rubrics, and scoring descriptors to reach an agreement on what should be tested and to reconcile their definitions and standards for the four components under the writing construct. However, the two raters are still different in years of teaching and learning experience, personality, grading leniency, and this was their first time working together as a team, all of which can be the possible reasons why not all individual variables have high and statistically significant correlations. In this section, we examined the level of construct validity for the writing task. There were four components, or variables, for the writing tasks: Language Control, Content Accuracy, Rhetoric Control, and Content Elaboration. The participants' average scores on each of the observed variables were imported into SPSS for correlation analysis. We chose the Person Product-Moment formula because those scores were composite in nature.

Table 18. Pearson Correlation Matrix for writing average: Observed Variables (N = 8)

	Language Cont Ave	Contt Accu Ave	Rhet Con Ave	Cont Elab Ave	
Person Correlation	0.61	.934**	.895**		**Cont Elab Ave**
Sig. (2-tailed)	0.108	<.001	0.003		
Person Correlation	0.667	.882*			**Rhet Con Ave**
Sig. (2-tailed)	0.071	0.012			
Person Correlation	0.348				**Contt Accu Ave**
Sig. (2-tailed)	0.398				
Person Correlation					**Language Cont Ave**

*. *p* < .05, 2-tailed
**. *p* < .01, 2-tailed

As the results in Table 18 show, Content Elaboration and Content Accuracy, among all the relationships between variables, showed a very high positive correlation at .934. It is also encouraging to see that the correlation was found to be statistically significant at the .01 level, which means there is a 99% chance that this correlation is not due to chance. Similarly, we observed that Content Accuracy and Rhetoric Control had a high correlation at .822, and Content Elaboration and Rhetoric Control had a high correlation at .895. Both correlations were found to be statistically significant, at the level of .05 and .01, respectively. It was also observed that Language Control displayed a positive correlational relationship with Content Accuracy, Rhetoric Control, and Content Elaboration, at .348, .667, and .610, respectively. However, none of the three correlations were statistically significant, which means the correlations could have occurred due to chance. Nevertheless, it is encouraging to see all the correlations are positive, and three of them were found to be statistically significant, suggesting that some of the variables were measuring the same underlying construct to a high degree, and some to a certain degree. This supports our previous argument based on the literature review that writing ability in this test consists of four components: language control, content accuracy, content elaboration, and rhetoric control.

4 Conclusion

In conclusion, the learning-oriented test model offers a valuable approach to assessing students' language proficiency, focusing on their learning progress and individual growth. Adopting a dynamic and student-centered perspective, this assessment method

emphasizes the significance of the learning process itself. However, it is essential to acknowledge the limitation of small sample sizes in some studies, which may impact the generalizability of findings. Future research should aim to address this limitation and incorporate larger and more diverse samples to enhance the validity and reliability of the assessment outcomes. Overall, the learning-oriented test model holds promise in promoting effective language learning and personalized educational practices, contributing to the advancement of language assessment in the educational context.

References

1. Abbott, R.D., Berninger, V.W., Fayol, M.: Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. J. Educ. Psychol. **102**(2), 281–298 (2010)
2. Alderson, C.: Assessing Reading. Cambridge University Press, Cambridge (2000)
3. Bachman, L.F., Palmer, A.S.: Language Testing in Practice. Oxford University Press, Oxford (1996)
4. Bazerman, C.: A relationship between reading and writing: the conversational model. Coll. Engl. **41**(6), 656–661 (1980)
5. Bennett, R.E., Deane, P., van Rijn, W., P.: From cognitive-domain theory to assessment practice. Educ. Psychol. **51**(1), 82–107 (2016)
6. Berninger, V.W., Abbott, R.D., Abbott, S.P., Graham, S., Richards, T.: Writing and reading: connections between language by hand and language by eye. J. Learn. Disabil. **35**(1), 39–56 (2002)
7. Borsboom, D., Mellenbergh, G.J., Van Heerden, J.: The concept of validity. Psychol. Rev. **111**(4), 1061 (2004)
8. Bridgeman, B., Carlson, S.: Survey of academic writing tasks required of graduate and undergraduate foreign students. ETS Res. Rep. Ser. **1983**(1), i–38 (1983)
9. Camp, R.: The place of portfolios in our changing views of writing assessment. In Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment, pp. 183–212 (1993)
10. Cao, Y., Chen, J., Zhang, M., Li, C.: Examining the writing processes in scenario-based assessment using regression trees. ETS Res. Rep. Ser. **2020**(1), 1–16 (2020)
11. Carson, J.E., Carrell, P.L., Silberstein, S., Kroll, B., Kuehn, P.A.: Reading-writing relationships in first and second language. TESOL Q. **24**(2), 245–266 (1990)
12. Cumming, A., Grant, L., Mulcahy-Ernt, P., Powers, D.E.: A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. Lang. Test. **21**(2), 107–145 (2004)
13. Davis, F.B.: Research in comprehension in reading. Read. Res. Q. **3**, 499–545 (1968)
14. Emig, J.: The composing processes of twelfth graders (1971)
15. Gebril, A., Plakans, L.: Investigating source use, discourse features, and process in integrated writing tests. In: Spaan Fellow Working Papers in Second or Foreign Language Assessment, , vol.7, no. 1, pp. 47–84 (2009)
16. Grabe, W.: Current developments in second language reading research. TESOL Q. **25**(3), 375–406 (1991)
17. Grabe, W., Zhang, C.: Reading-writing relationships in first and second language academic literacy development. Lang. Teach. **49**(3), 339–355 (2016)
18. Hamid, M.O., Hardy, I., Reyes, V.: Test-takers' perspectives on a global test of English: questions of fairness, justice and validity. Lang Test Asia **9**, 16 (2019)
19. Hayes, J. R.: Understanding Cognition and Affect in Writing. Perspectives on writing: Research, theory, and practice, vol. 6 (2000)

20. Horowitz, D.: Essay examination prompts and the teaching of academic writing. Engl. Specif. Purp. **5**(2), 107–120 (1986)
21. Kim, A.Y.: Investigating second language reading components: Reading for different types of meaning (2009)
22. Lumley, T.: The notion of subskills in reading comprehension tests: an EAP example. Lang. Test. **10**(3), 211–234 (1993)
23. Munby, J.: A problem-solving approach to the development of reading comprehension skills. In: Presentation at the University Teachers of English in Israel (UTELI) Regional Meeting, Jerusalem, January, vol. 25 (1978)
24. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet Rasch measurement: Part I. J. Appl. Meas. **4**(4), 386–422 (2003)
25. O'hare, F.: Sentence Combining: Improving Student Writing without Formal Grammar Instruction. NCTE Committee on Research Report Series, No. 15 (1973)
26. Plakans, L., Gebril, A.: A close investigation into source use in integrated second language writing tasks. Assess. Writ. **17**(1), 18–34 (2012)
27. Plakans, L., Gebril, A.: Using multiple texts in an integrated writing assessment: source text use as a predictor of score. J. Second. Lang. Writ. **22**(3), 217–230 (2013)
28. Purpura, J.E.: The development and construct validation of an instrument designed to investigate the cognitive background characteristics of test-takers. In: Kunnan, A.J. (ed.), Validation in Language Assessment, pp. 111–139. Mahwah, NY: Lawrence Erlbaum Associates, Inc (1998)
29. Purpura, J.: Assessing grammar. Cambridge University Press, Cambridge, UK (2004)
30. Purpura, J.E.: A Rationale for Using a SBA to Measure Competency-Based, Situated S_FL Proficiency. Jan 3 final version manuscript (2021)
31. Purpura, J.E.: Class notes from Second Language Assessment (2022)
32. Raimes, A.: Out of the woods: emerging traditions in the teaching of writing. TESOL Q. **25**(3), 407–430 (1991)
33. Schoonen, R.: How language ability is assessed. Handb. Res. Second Lang. Teach. Learn. **2**, 701–716 (2011)
34. Schoonen, R.: Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. Read. Writ. **32**(3), 511–535 (2019)
35. Selinker, L., Todd-Trimble, M., Trimble, L.: Rhetorical function-shifts in EST discourse. TESOL Q. **12**, 311–320 (1978)
36. Selzer, J.: The composing processes of an engineer. Coll. Compos. Commun. **34**(2), 178–187 (1983)
37. Shanahan, L.E.: Reading and writing multimodal texts through information and communication technologies (2006)
38. Weigle, S.C.: Assessing Writing. Cambridge University Press, Cambridge (2002)
39. Weigle, S.C.: Integrating reading and writing in a competency test for non-native speakers of English. Assess. Writ. **9**(1), 27–55 (2004)
40. Wiersma, W., & Jurs, S. G.: Research Methods in Education: An Introduction (9th ed.) (2009)
41. White, R.V.: New Ways in Teaching Writing. New Ways in TESOL Series: Innovative Classroom Techniques (1995)
42. Zamel, V.: Re-evaluating sentence-combining practice. Tesol Q. 81–90 (1980)
43. Carr, N.: Designing and Analyzing Language Tests. Oxford University Press, Oxford, UK (2011)