Designing and Researching an Intertextual Reading-Into-Writing Summary Task

Nathaniel Owen, Oxford University Press, United Kingdom

The European Conference on Language Learning 2023
Official Conference Proceedings

Abstract

This paper reports on the design and evaluation of an innovative intertextual reading-into-writing task for use in academic admissions. Existing tests of English used for university admission avoid intertextual reading (Weir and Chan, 2019) and do not ask test takers to synthesize information from multiple texts into a single piece of writing. Therefore, there is a divergence between the design of language tests for university entrance and subsequent 'academic writing' required at university. We designed a mediation-focused Summary task which requires test takers to read two texts on the same topic (a total of 300 words) and to summarize the information using their own words (up to a maximum of 100 words). Seven trained judges provided CEFR ratings for 48 internally benchmarked test taker scripts across four Summary tasks (n=24) and four Essay tasks (n=24). Data were analyzed using many-facet Rasch analysis to investigate task, judge and rating scale performance. We also analyzed the level of agreement with internal benchmarking. We found a strong level of agreement between internal and external expert assessors and that the task is highly effective for distinguishing B2 from C1-level performances. Assessors were able to score responses using an analytic rating scale incorporating source use across four individual components at similar levels of reliability to a more traditional Essay task. Test takers' lack of familiarity with the task design means the introduction of such tasks will have a significant washback effect.

Keywords: Assessment, Summarizing, Integrated-Skills

iafor
The International Academic Forum
www.iafor.org

1. Introduction

Language testing has long been predicated on a ‘four skills’ approach of speaking, reading, writing and listening (Lado, 1961). Tests of English predicated on skills-based modules may treat these skills as unnaturally isolated, whereas in the domains for which the tests were created, these skills are often combined to complete specific tasks (Gebril & Plakans, 2014; Plakans & Gebril, 2012; Yu, 2013) which “more closely resemble the kinds of language use tasks examinees are expected to encounter in everyday life” (Sawaki, Stricker & Oranje, 2009, p. 5). When considering higher education more specifically, universities have emphasized the importance not only of text comprehension, but of other crucial skills such as summarizing and synthesizing information and reading complex texts without instruction or guidance (University of California, 2002). The importance of integrated skills is also evident in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), which include the concepts of ‘interaction’ and ‘mediation’, both of which emphasize combining skills for describing overall language ability. As a result, Oxford University Press has sought to develop an integrated intertextual reading-into-writing task, developed from the ground up against expanded CEFR guidelines, for use in university entrance and professional contexts. This paper reports on the initial development and evaluation of this task.

2. Literature

Given the renewed focus on mediation in the CEFR Companion Volume (Council of Europe, 2020), this formed the starting point for developing the task.

2.1. Mediation in the CEFR

‘Mediation’ typically refers to ‘indirect conveyance or communication through an intermediary.’ Within the CEFR Companion Volume (2020), the concept of mediation is “a social and cultural process of creating conditions for communication and cooperation, facing and hopefully defusing any delicate situations and tensions that may arise” (Council of Europe, 2020, p. 106). This includes both cross-linguistic mediation and mediation within a target language. In both conceptions, it is chiefly concerned with facilitating the communicative needs of others. It is also significant that mediation often happens across modalities, so that written output may for example involve processing and relaying the message of a spoken text or synthesis of multiple sources.

2.2. An Intertextual Reading-Into-Writing Construct

The purpose of developing an integrated reading into writing task is to better represent higher order processing in reading and writing that is a cornerstone of academic proficiency. When students are required to read for writing, they will use appropriate reading strategies to construct elaborate models of text structure using higher order processes to construct textual and intertextual representations which allow them to select, evaluate and use information according to the writing purpose (Weigle, Yang, & Montee, 2013). In academic settings, students are required to read *multiple* sources and integrate information into extended pieces of writing using their own words. Although integrated-skills tasks are not new, existing tests of English used for university admission or professional purposes have completely eschewed intertextual reading (Owen, 2016; Weir and Chan, 2019) and do not ask test takers to synthesize information from multiple texts into a single piece of writing. Asking test takers to write a *summary* of two texts requires test takers to read and differentiate main from supporting information in two (or more) texts, then *synthesize* that information into a new text for a specified audience.

2.2.1. Task Requirements: Discourse Synthesis in Summary Writing Tasks

Discourse synthesis (Nelson and King, 2022) refers to operations such as organizing, selecting, and connecting content from multiple sources on the same topic. Therefore, marking criteria should examine content transformation and degree of source use rather than just appropriation. This is crucial to successful summary writing. Higher scores should be awarded for making explicit links across sources, especially where such links may only be inferred. Summary tasks could employ an upper word limit rather than a minimum word limit traditionally associated with second language writing tests to ensure idea selection and transformation rather than just reproduction. This is crucial to a successful summary task, as Crossley et al. (2023) found that text length proved to be the strongest predictor of test taker performance in such tasks.

Given increased cognitive demands placed on L2 writers in processing multiple texts with a subsequent writing requirement, sufficient planning time is a key element of successfully creating a summary task (Leijten et al, 2019). Hyland (2009) notes that L2 writers tend to plan less than L1 writers, encountering more difficulty in setting goals and generating text. Shaw and Weir (2007) also note that novice writers plan very little and focus on generating content from within remembered resources from the topic or genre. However, skilled writing entails a heightened awareness of task purpose. A strong reading ability, including the ability to identify reading purpose, has been linked to characteristics of students' synthesis of input texts such as successful organization of ideas (Spivey, 1988).

2.2.2. Source Text Genre in Summary Writing Tasks

Explicitly stating genre is crucial, as Li (2014) notes that source text genre has a significant impact on test taker performance in summary tasks. Narrative and expository texts pose different challenges and elicit different strategies from students. Li found that students perform better when summarizing an expository text compared to a narrative text. Students experience greater difficulty identifying main ideas and creating a thesis statement for narrative text. This is because expository texts contain more explicit topic sentences and hierarchical structures compared to narrative texts with linear plot structures.

2.2.3. Designing Rating Scales for Assessing Test Taker Performance in Summary Writing Tasks

Productive components of language tests (speaking and writing) are scored using rating scales. Rating scales can be analytic or holistic. Holistic scales are those in which a single score is awarded to a sample of writing (Hamp-Lyons, 1991) and analytic rating scales those in which multiple scores are awarded to the same sample, each of which represents an aspect of the construct as identified by the test developer. For a fuller review of the advantages and disadvantages of analytic and holistic rating scales, see Barkaoui (2011). Analytic scales tend to be preferred for integrated reading-into-writing tasks (Lestari & Brunfaut, 2023; Lestari & Ho, 2023). Developers must decide whether source use is a separate scale or whether this is integrated into descriptors for other rating scale components. Lestari and Ho (2023) compared a scale with a separate criterion for reading/source use with one integrating reading aspects into other criteria. Analysis indicated both scales functioned well, but the separate criterion was slightly clearer to raters. In contrast, Leijten et al (2019) recommend avoiding source use in language criteria to prevent assessor confusion. When investigating the performance of analytic scales for reading-into-writing tasks, dimensionality analysis is therefore paramount for analyzing the performance of the different scales or whether assessors are unintentionally focusing on one aspect of language (Knoch et al, 2020; Leijten et al, 2019).

2.3. Research Questions

The discussion of the literature in Section 2.2 reveals that there are many ways in which the validity of a summary writing task can be investigated. However, as this is a new task and rating scale, we restricted our initial investigations to the functioning of the rating scale and the suitability of the task for its intended purposes. Therefore, the following research questions were devised to investigate task and rating scale performance:

RQ1: To what extent does test taker performance in the summary task align with other measures of writing proficiency?

RQ2: To what extent do each of the analytic scale components provide unique information?

3. Methodology

To address the research questions, the study employed a quantitative research design, analyzing numerical data collected from assessors to make judgements about task suitability and rating scale performance. The overall research design is presented in Table 1:

Data Collection	**Data Analysis**	**Research question**
• Rating of 24 summary task performances by seven assessors using rating scale aligned to the CEFR	• Many-facet Rasch measurement (rating scale model) of Essay and Summary data • Correlation analysis • Separate Rasch analysis of Essay and Summary data	RQ1. How does test taker performance in the summary task align with other measures of writing proficiency?
• Rating of 24 Essay task performances by seven assessors aligned to the CEFR		
	• Many-facet Rasch measurement (partial credit model) of Essay and Summary data • Correlations between components of the rating scale	RQ2. Does each of the components provide unique information?

Table 1. Research design

3.1. Assessors

Seven assessors (five male, two female) participated in this study. All were English first language (L1) speakers, while all held English language teaching qualifications. Three hold PhDs in language assessment while the remaining four possess extensive experience in English language teaching, item writing and materials development. All participants possess at least a BA. Three participants had experience working as professional assessors for high-stakes standardized English language tests, including experience with analytic rating scales of the kind used in this research. The assessors were recruited through professional networks via email and provided signed consent forms.

3.2. Research Instruments

Having considered the lessons from both the CEFR and recent literature, Oxford University Press designed a summary task as represented by the example in Figure 1:

You have 20 minutes to write a summary. Write 80–100 words.

You have been learning about psychology in a college class. Your tutor has now asked you to read about research in psychology and write a summary of the main ideas to share with your class.

Read the two texts below and write one paragraph using full sentences, combining the information given in the texts. Use your own words where possible. Your summary should provide the reader with enough information to understand the main ideas in the texts.

Do **NOT** write more than 100 words. Write your summary.

Psychology textbook extract	**Psychology lecture transcript**
In theory, when researchers conduct research, their experiments are not biased and provide valid results. In practice, the research results can be negatively affected in various ways. One common cause of bias is the researcher themselves. For example, a researcher can make mistakes when recording results. This is referred to as the 'experimenter effect' as it is the experimenter that affects the outcome of the experiment, reducing confidence in the experiment's result. There are two main kinds of experimenter effects. First, let's turn to non-interactional effects. These effects are found in research that does not require the researcher to interact with the research subjects, for example where the researcher does not record accurately what they have observed, known as the Observer Effect. Another example is where the researcher interprets the evidence from an experiment incorrectly, known as the Interpreter Effect. Less common are Intentional Effects, where researchers do not report the research results accurately on purpose.	"Back in the late sixties, psychologist Robert Rosenthal conducted an experiment into 'Interactional Effects', that is, ones that involve the researcher interacting with the subject. Two teams of researchers were set up, each given a maze containing rats to observe. One team was told they had intelligent rats; the other team were told they had unintelligent rats. The intelligent rats solved the maze well while the unintelligent rats didn't. What makes this surprising is that the rats, in fact, were all equally intelligent! This is a classic case of 'expectancy effect', where the researcher unconsciously influences the subject to act in a way the researcher wants them to, making the research less valid. It's believed that the 'intelligent' rats did better because the researchers with the 'intelligent' rats treated them better than those with 'unintelligent' rats."

GLOSSARY

experimenter: a researcher
maze: a system of paths with walls designed so that it is difficult to find your way through.
subject: a person or animal that a researcher collects information on.
unconscious: if you are unconscious of something when you are doing it, you are not aware you are doing it.
unintelligent: not intelligent

Figure 1. Example of written summary task for the Oxford Test of English Advanced

The task contains instructions, two texts and a glossary. The time limit for the task is 20 minutes, including both reading and writing time. The task rubric specifies the context and topic, in this case psychology. Test takers are specifically asked to write a single paragraph in full sentences to avoid the impulse to write bullet points. The instructions specifically call on the test taker to identify the main ideas in the texts. The word limit for the task is 100 words to avoid the issue of test takes trying to write too much. There is a multimodal aspect to the task, as one text is an extract from a textbook, and is a lecture transcript. There is a glossary which identifies and provides short definitions of low- frequency lexis. However, subject-specific vocabulary is *not* defined, as this would overly assist the test takers with the task.

3.3. Rating Scales for the Summary Task

To score test taker responses, assessors use an analytic rating scale. The rating scale contains four components: task fulfilment, organization, grammar and lexis. Source use is integrated across all four criteria, rather than having an independent scale. Scores range from 'Below B1' to 'Above C1' in half-band increments. Each test taker therefore receives four scores, one each for task completion, organization, grammar and lexis (maximum score = 28). The full rating scale can be viewed in Appendix A.

3.4. Assessor Training

Assessors were provided with a series of materials, including the summary task specification, summary rating scale (Appendix A), additional guidance on using the criteria as well as CEFR materials on mediation. Assessors were also provided with twelve samples of test taker responses to two different summary tasks (six samples per summary task). Assessors were asked to familiarize themselves with the materials, then use the rating scale to score the twelve samples. Assessors entered their scores into an Excel spreadsheet which were then returned to Oxford University Press. These scores were used as the basis for a synchronous online training webinar. The webinar lasted approximately two hours. Assessors were able to ask questions about the rating scale. Three samples from the training were selected based on the assessors scores – two which had strong disagreement, and one which had strong agreement. These were used as the basis of discussion among participants to come to a shared understanding of the criteria.

3.5. Main Data Collection

Upon the conclusion of the webinar, assessors were given access to the samples used for the main data collection. Four summary tasks were used in this research. Six test takers provided responses to each summary, resulting in a total of 24 samples of test taker writing. Samples were selected from pretesting conducted in 2022. The samples had all been scored internally and were chosen to represent a range of scores from B1-C1 of the CEFR. All 24 samples of writing were marked by all seven assessors, who gave four scores per sample (task fulfilment, organization, grammar and lexis). The assessors did not have access to the scores initially awarded to the samples during pretesting. The same data collection procedure was performed for 24 test takers who completed Essay tasks. Of these 24 test takers, 17 also completed the summary tasks. These seventeen test takers provided the basis for addressing Research question 1, by comparing their performance on the Essay task with the Summary task. The Essay task also has an analytic rating scale with four components (task fulfilment, organization, grammar and lexis). However, the content of the scales is different, due to differing task requirements. Data collection took place in February – March 2023.

3.6. Methods of Data Analysis

Assessors' ratings were analyzed using many-facet Rasch measurement (MFRM) within the programs FACETS v3.84 (Linacre, 2023a) WINSTEPS v5.2.4.0 (Linacre, 2023b). MFRM is a variant of Rasch measurement used when data is reported on a rating scale by independent judges. To address Research question 1, a five-facet model was adopted (assessors, test-takers, task (Essay/Summary), test, component). Because each test had a unique cohort with no overlap (no test taker took more than one test), the test facet was used as a dummy variable (anchored at zero) to link the dataset and avoid the emergence of subsets. The research adopted a variation of MFRM called the rating scale model (Wright and Masters, 1982). This model specifies the probability, Pnij, that person *n* of ability measure Bn is observed in category *j* of a rating scale *F* specific to item *i* of difficulty measure Di as opposed to the probability Pni(j-1) of being observed in category (j-1):

$$\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_{ij}$$

In this variant, the rating scale structure {Fij} becomes specific to item *i*, although difficulty, ability and assessor leniency measures are still plotted on the same scale in the output. Additionally for both Research questions 1 and 2, the rating scale model (Andrich, 1978) was also employed to further explore the individual components as independent criteria. In the rating scale model, the partial credit model treats individual rating scale criteria as having their own scale structure. This is expressed as:

$$\log_e(P_{nij} / P_{ni(j-1)}) = B_n - D_{gi} - F_{gj}$$

The subscript 'g' in the rating scale model specifies the group of items to which item *i* belongs and identifies the rating scale structure that belongs to the group (Linacre, 2023c, p. 3). This allows for scrutiny of the scale performance for each component of the rating scale. This model assumes equal threshold parameters for each component and so is appropriate for analyzing the functioning of an analytic scale (McNamara, Knoch & Fan, 2019). The same analysis was performed on the Essay data. Correlations were also calculated for the test takers who had taken both the Summary and the Essay task.

4. Findings

RQ1: To what extent does test taker performance in the summary task align with other measures of writing proficiency?

To address Research question 1, we directly compared the performance of test takers in the Summary task to their performance in the Essay task. This was initially addressed by examining the MFRM output from the partial credit model (Table 2). This model incorporated five facets, of which four are reported below. The fifth facet was rating scale component, but as the content of the scales are different for the Essay and Summary, this facet is not reported here.

	Test taker	Assessor	Pretest	Task (Essay/Summary)
Measure				
M	-0.07	0.00	0.00*	0.00
SD (pop)	1.02	0.68	0.00*	0.43
Average SE	0.17	0.08	0.06	0.04
n	31**	7	4	2
InfitMS				
M	0.94	0.99	1.00	1.00
SD (pop)	0.38	0.23	0.14	0.14
OutfitMS				
M	0.96	1.02	1.03	1.03
SD (pop)	0.41	0.28	0.16	0.18
Separation statistics				
Ratio	6.81	8.69	0.00	10.25
Strata	9.41	11.93	0.33	14.00

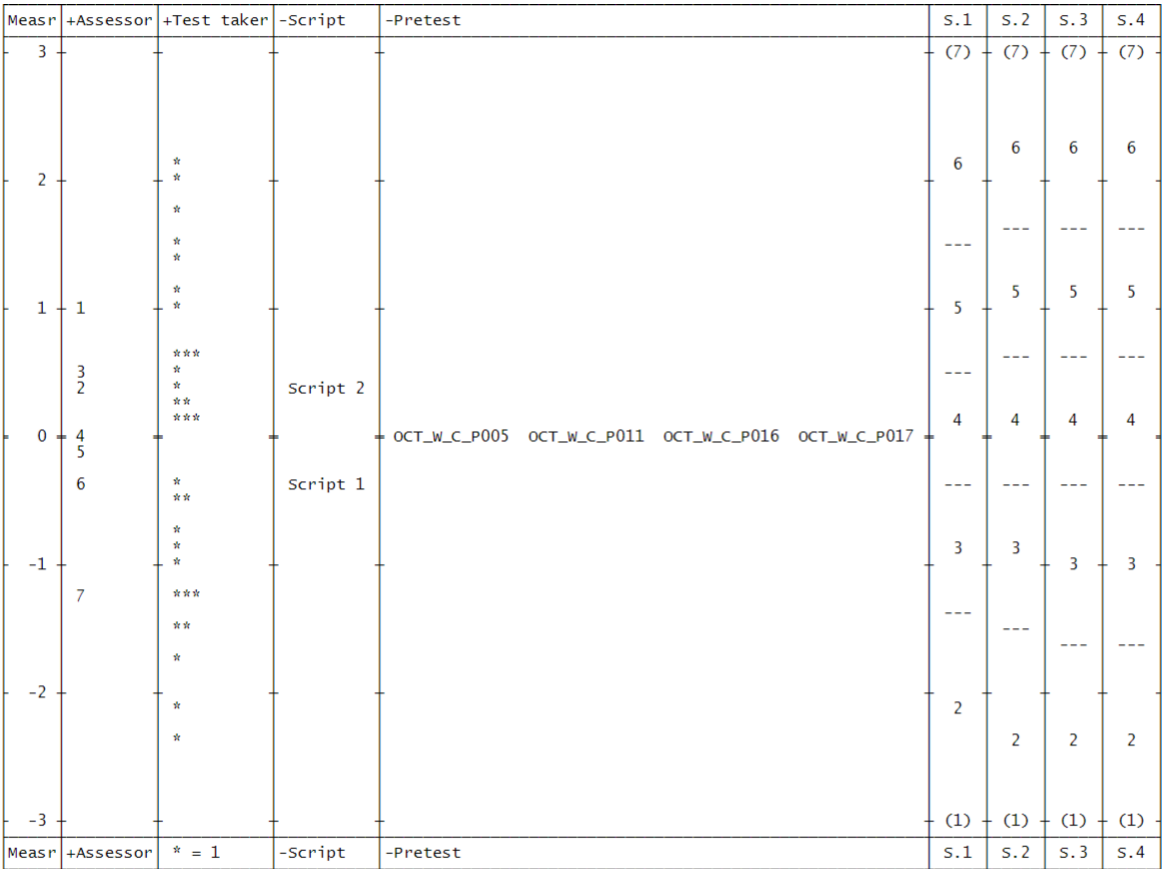
Reliability	.98	.99	.00	.99
Fixed X^2	1513.1***	503.3***	0.00	212.3***
df	30	6	3	1

*Pretest anchored at zero (dummy variable). Assessor facet mean set to zero. Test taker facet allowed to float.
**24 test takers each took an Essay and Summary task in total; only 17 test takers took both tasks.
***$p < .001$.

Table 2. Summary statistics of three facets (MFRM rating scale model)

Table 2 provides details of the performance of the three facets. This is interpreted through fit statistics: infit and outfit mean square (MS) values. These have an expected value of 1 (Linacre, 2023d), with deviations above and below this indicating that the collected data are unproductive for measurement. McNamara (1996) states an acceptable range for this statistic is between 0.7-1.3, although Lunz, Wright and Linacre (1990) suggest a less strict criteria of 0.6 to 1.5 for fit statistics. The figures for test taker, assessor and task facets were all close to or exactly 1, indicating that data for the Summary and Essay tasks both meet expectations for productive measurement. The fixed Chi-square statistic tests the hypothesis that elements of each facet share consistent patterns and are thus of the same ‘type’. The highly significant results suggest that there are different types in every facet. For example, the separation statistic (6.81) indicates that a rating scale with seven levels is appropriate for describing the performances in the test which were selected to represent different CEFR bands. The large figure for task (10.25) indicates that the score patterns for each task are sufficiently different from each other and that the tasks are measuring different skills.

The overall outcome of the MFRM analysis is represented in Figure 2 below, which shows the vertical ruler, or FACET map, for all four facets in the model (with pretest anchored at zero).



```
|Measr|+Assessor|+Test taker|-Script    |-Pretest                                                  | S.1 | S.2 | S.3 | S.4 |
|-----+---------+-----------+-----------+----------------------------------------------------------+-----+-----+-----+-----|
+   3 +         +           +           +                                                          + (7) + (7) + (7) + (7) +
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |  6  |  6  |  6  |
|     |         |           |           |                                                          |     |     |     |     |
|     |         | *         |           |                                                          |  6  |     |     |     |
+   2 +         + *         +           +                                                          +     +     +     +     +
|     |         |           |           |                                                          |     |     |     |     |
|     |         | *         |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     | --- | --- | --- |
|     |         | *         |           |                                                          | --- |     |     |     |
|     |         | *         |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         | *         |           |                                                          |     |  5  |  5  |  5  |
+   1 + 1       + *         +           +                                                          +  5  +     +     +     +
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         | ***       |           |                                                          |     | --- | --- | --- |
|     | 3       | *         |           |                                                          | --- |     |     |     |
|     | 2       | *         | Script 2  |                                                          |     |     |     |     |
|     |         | **        |           |                                                          |     |     |     |     |
|     |         | ***       |           |                                                          |  4  |  4  |  4  |  4  |
*   0 * 4       *           *           * OCT_W_C_P005  OCT_W_C_P011  OCT_W_C_P016  OCT_W_C_P017 *     *     *     *     *
|     | 5       |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     | 6       | *         | Script 1  |                                                          | --- | --- | --- | --- |
|     |         | **        |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         | *         |           |                                                          |     |     |     |     |
|     |         | *         |           |                                                          |  3  |  3  |     |     |
+  -1 +         + *         +           +                                                          +     +     +  3  +  3  +
|     |         |           |           |                                                          |     |     |     |     |
|     | 7       | ***       |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          | --- |     |     |     |
|     |         | **        |           |                                                          |     | --- |     |     |
|     |         |           |           |                                                          |     |     | --- | --- |
|     |         | *         |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
+  -2 +         +           +           +                                                          +     +     +     +     +
|     |         | *         |           |                                                          |  2  |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         | *         |           |                                                          |     |  2  |  2  |  2  |
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
|     |         |           |           |                                                          |     |     |     |     |
+  -3 +         +           +           +                                                          + (1) + (1) + (1) + (1) +
|-----+---------+-----------+-----------+----------------------------------------------------------+-----+-----+-----+-----|
|Measr|+Assessor| * = 1     |-Script    |-Pretest                                                  | S.1 | S.2 | S.3 | S.4 |
```

Figure 2. All FACET vertical rulers from rating scale model analysis

Facets 1 (assessor leniency) is oriented such that positive logits indicate greater amounts of the construct (leniency). That is, the higher the score, the more lenient the assessor. Facet 2 displays the weakest test takers are towards the top of the measurement scale, with stronger writers at the bottom. Facet 3 (Essay/Summary difficulty) is oriented such that higher scores indicate greater difficulty. The assessors are mostly clustered together between -1 and +1 logits, showing that most were not excessively harsh or lenient, except for assessor 7, who appeared harsher than other assessors by almost a full logit (assessor 7 = -1.30, assessor 6 = -0.41). The test-takers in column 3 are dispersed from -2.40 to +2.18 logits, indicating that the approach to test taker selection was effective in identifying a range of test taker abilities for the research and represents evidence that both Essay and Summary tasks are appropriate for assessing language ability at different levels of the CEFR. Task difficulty showed that the Summary task was almost a full logit more challenging than the Essay task (Essay = 0.43; Summary = -0.43). The full measurement reports for assessor, test taker and task facets can be seen in Appendices B, C and D respectively.

To look at the task data in more depth, we compared the performances of test takers in the Summary and Essay tasks directly. As noted in Section 3.5, the samples used in the data analysis were all collected during pretesting and had been double marked internally. These scores were used as the basis for sample selection. This allowed us to correlate the internal scores with the fair mean scores derived from the FACETS analysis of the ratings of the seven participants. The fair mean is a Rasch measure to raw score conversion, producing an average rating for each element that is standardized against average values of the facet (Linacre, 2023d, p. 211). The data is depicted in Figure 3 and Table 3 below.

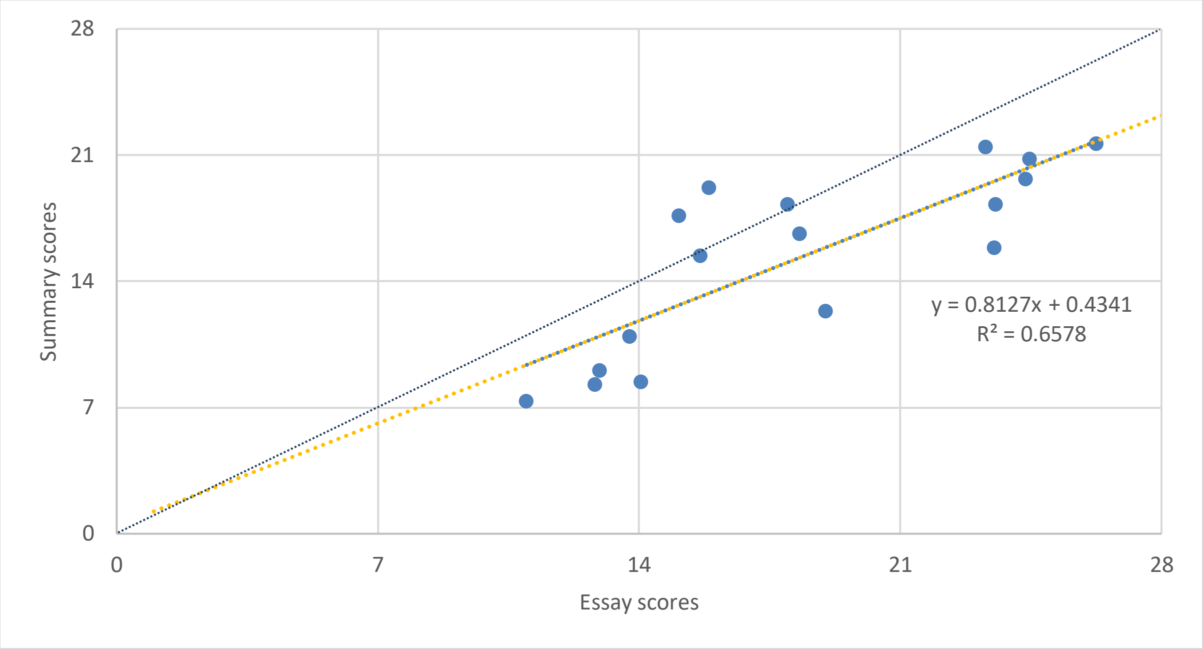


Figure 3. Comparison of Essay and Summary scores

Correlations		**Agreement (Internal vs. External)**		
Essay	0.90	Essay	19	24
Summary	0.94	Summary	18	24
Overall	0.91	Overall	37	48
				0.77

Table 3. Correlations and agreement between internal and assessor scores

Figure 3 shows the relationship between raw scores awarded to the seventeen test takers who completed both a Summary and Essay task. They show a strong association, with approximately 66 percent shared variance. The blue line is the ideal line, in which participants would receive the same score for both the Essay and the Summary. However, only three test takers are above this line, with the majority below it. This indicates that test takers generally receive slightly lower scores for the Summary than the Essay, indicating that they found this task more challenging. Table 3 shows correlations between the scores awarded by the seven assessors and the scores awarded to the samples from pretesting and the level of agreement between the CEFR bands awarded during pretesting and CEFR bands awarded by the seven assessors. 18 out of the 24 Summary samples were awarded the same CEFR band by the assessors as awarded during internal pretesting. This was consistent with the finding for the Essay task. Correlations are high (above .9), indicating strong agreement between internal markers and the assessors. The overall level of agreement for awarding CEFR bands to test takers was .77.

Figure 4 below shows the fair mean scores for each component derived from the partial credit model. This model was used to analyze each component separately to obtain the fair means for each test taker for each component. Scores for the Essay and Summary tasks were then plotted against each other.

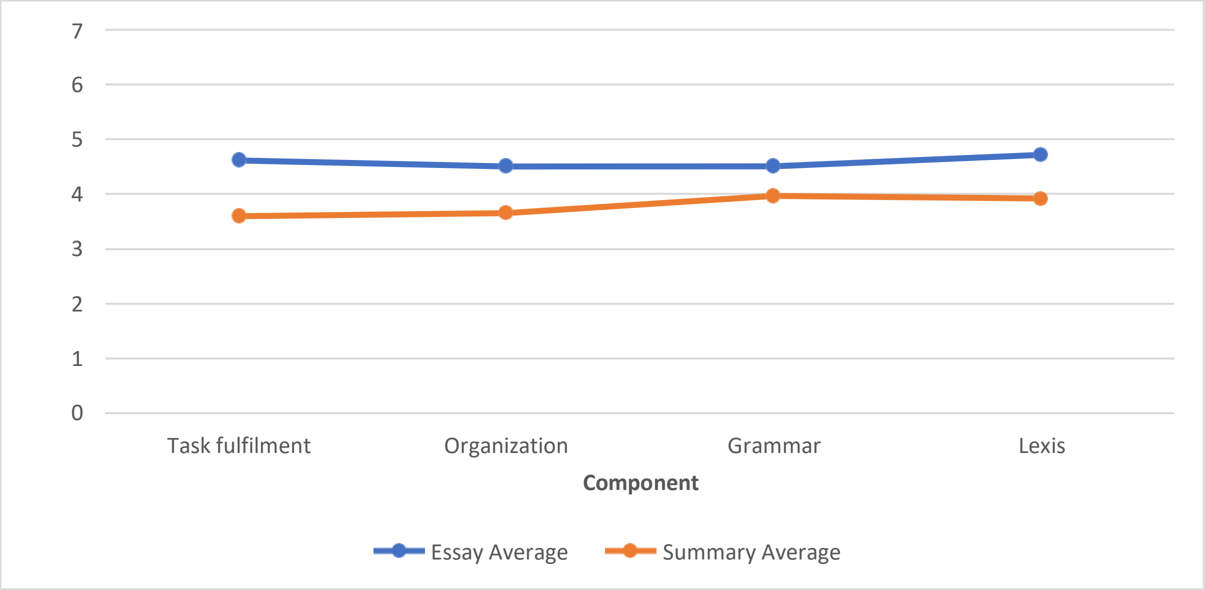
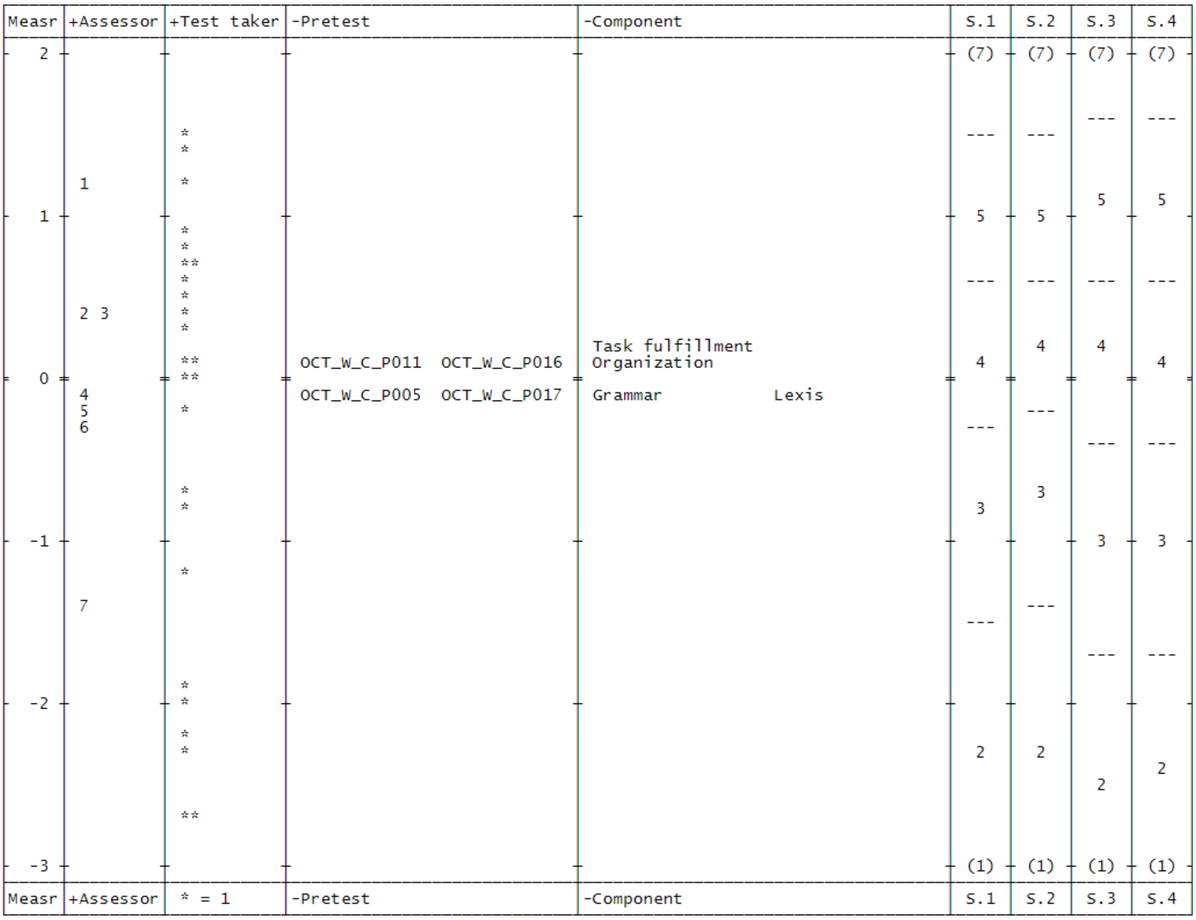


Figure 4. Fair mean average component scores for Essay and Summary tasks

Scores for the summary responses were uniformly lower than Essay scores but were closer for grammar and lexis than they were for task fulfilment and organization. This suggests that participants (both test takers and assessors) were less familiar with the Summary task design and that test takers were rewarded for their language use (grammar and lexis) but marked down for task and organization due to this lack of familiarity potentially affecting their task performance. This difference then impacted their overall scores as seen in Figure 3.

***RQ2*: To what extent do each of the analytic scale components provide unique information?**

Research question 2 was addressed in two ways. First, the rating scale model was used to analyze the Summary data independently of the Essay data. Secondly, Essay and Summary data were analyzed using the partial credit model in WINSTEPS to investigate the multidimensionality of the four components of the rating scales used in the Summary task. The fair means depicted in Figure 4 already suggest that different components of the rating scales may have different difficulty levels. In the FACETS analysis, the pretests were unanchored to explore task difficulty in more detail. Dummy test takers were anchored at zero and weighted at .0001 to link the dataset and eliminate the problem of disjoint subsets. The outcome of the analysis is shown in Figure 5:



```
--------------------------------------------------------------------------------------------------------------------------
|Measr|+Assessor|+Test taker|-Pretest                     |-Component                           | S.1 | S.2 | S.3 | S.4 |
--------------------------------------------------------------------------------------------------------------------------
+   2 +         +           +                             +                                     + (7) + (7) + (7) + (7) +
|     |         |           |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     | --- | --- |
|     |         | *         |                             |                                     | --- | --- |     |     |
|     |         | *         |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     | 1       | *         |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |  5  |  5  |
+   1 +         +           +                             +                                     +  5  +  5  +     +     +
|     |         | *         |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |     |     |     |     |
|     |         | **        |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     | --- | --- | --- | --- |
|     |         | *         |                             |                                     |     |     |     |     |
|     | 2 3     | *         |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |     |     |     |     |
|     |         |           |                             | Task fulfillment                    |     |  4  |  4  |     |
|     |         | **        | OCT_W_C_P011  OCT_W_C_P016  | Organization                        |  4  |     |     |  4  |
+   0 +         + **        +                             +                                     +     +     +     +     +
|     | 4       |           | OCT_W_C_P005  OCT_W_C_P017  | Grammar           Lexis             |     |     |     |     |
|     | 5       | *         |                             |                                     |     | --- |     |     |
|     | 6       |           |                             |                                     | --- |     |     |     |
|     |         |           |                             |                                     |     |     | --- | --- |
|     |         |           |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |     |  3  |     |     |
|     |         | *         |                             |                                     |  3  |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
+  -1 +         +           +                             +                                     +     +     +  3  +  3  +
|     |         |           |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     | 7       |           |                             |                                     |     | --- |     |     |
|     |         |           |                             |                                     | --- |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     | --- | --- |
|     |         |           |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |     |     |     |     |
+  -2 +         + *         +                             +                                     +     +     +     +     +
|     |         |           |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |     |     |     |     |
|     |         | *         |                             |                                     |  2  |  2  |     |     |
|     |         |           |                             |                                     |     |     |     |  2  |
|     |         |           |                             |                                     |     |     |  2  |     |
|     |         |           |                             |                                     |     |     |     |     |
|     |         | **        |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
|     |         |           |                             |                                     |     |     |     |     |
+  -3 +         +           +                             +                                     + (1) + (1) + (1) + (1) +
--------------------------------------------------------------------------------------------------------------------------
|Measr|+Assessor| * = 1     |-Pretest                     |-Component                           | S.1 | S.2 | S.3 | S.4 |
--------------------------------------------------------------------------------------------------------------------------
```

Figure 5. FACET vertical rulers from rating scale model analysis (Summary data only)

Figure 5 displays the relationship between test-taker ability, assessor leniency and rating scale component difficulty with the individual components modelled with their own scale structure. S1-S4 respectively represent task fulfilment, organization, grammar and lexis. Category thresholds are marked by the three dashes between each score within the rating scales. The data confirms that task fulfilment and organization were slightly more challenging than grammar or lexis. Scale data can be seen in Appendix E. Task fulfilment and organization recorded logit values of .17 and .06 respectively, while grammar and lexis recorded -.14 and -.09 respectively. However, all fit statistics (infitMS and outfitMS) were close to 1, suggesting that the data meet the expectations of the Rasch model for rating scale development. Scale discrimination was investigated further by performing separate three-facet analysis (test taker, assessor, pretest) on each of the four components of the rating scale. The outcome of this analysis is presented in Table 4.

Component	**Range of test taker ability (Logits)**		**Difference**
	Low	**High**	
Task fulfilment	-4.20	1.43	5.63
Organization	-3.25	1.46	4.71
Grammar	-3.05	1.71	4.76
Lexis	-3.58	1.62	5.20

Table 4. Scale component discrimination

The data confirms the initial finding that task fulfilment and organization proved to be more challenging for test takers than grammar or lexis (which recorded higher average ability estimates), and a slight truncation (narrowing) for organization, suggesting fewer score bands are being used for this component. To investigate this, we explored the monotonicity of the scale. This means that the average test taker ability should *increase* with each category of the rating scale. This was investigated by exploring the Rasch-Andrich thresholds for each of the four components, the outcome of which can be seen in Table 5.

	Task fulfilment		**Organization**		**Grammar**		**Lexis**	
Score	**Measure**	**SE**	**Measure**	**SE**	**Measure**	**SE**	**Measure**	**SE**
1	*	*	*	*	*	*	*	*
2	-3.03	.31	-3.25	.34	-3.02	.37	-2.85	.35
3	-1.39	.22	-1.26	.22	-1.78	.25	-1.78	.25
4	.21	.20	.29	.20	-.15	.20	-.08	.21
5	.25	.21	.21	.20	.72	.20	.52	.20
6	1.31	.25	1.45	.25	1.28	.23	1.33	.23
7	2.66	.47	2.56	.43	2.95	.43	2.85	.43
*Bottom category; no threshold below this level (scores of zero were not used by assessors)								

Table 5. Rasch-Andrich thresholds of the four components of the Summary rating scale

We can see for organization, scores of 4 were under-utilized relative to scores of 2, 3, 5 and 6, as the scale is not monotonic (the ability measure for Band 5 is lower than the ability measure for Band 4) suggesting that either this ability band was under-represented in the relatively small sample size of 24 test takers. To eliminate the possibility this was caused by systematic misuse of the organization criterion, we then compared the performance of each component of the rating scale across the Essay and Summary tasks. To do this, we examined the average Rasch values for each assessor in each component for the Essay and Summary tasks. This data was then plotted in a scatter graph in Python v.3.11.4, using the standard errors for each Rasch value to calculate 95 percent confidence intervals for a regression line. The outcome of this analysis can be seen in Table 6 and Figure 6.

NAME	**Component**	**Essay Rasch Value**	**Essay SE**	**Summary Rasch Value**	**Summary SE**
ASSESSOR 1*	Task	-1.23	0.26	-0.10	0.21
ASSESSOR 1*	Organization	-0.94	0.26	-0.19	0.21
ASSESSOR 1	Grammar	-0.71	0.25	-0.61	0.22
ASSESSOR 1*	Lexis	-1.57	0.26	-0.63	0.22
ASSESSOR 2	Task	-0.43	0.24	-0.92	0.22
ASSESSOR 2	Organization	-0.42	0.25	-1.01	0.22
ASSESSOR 2*	Grammar	-0.02	0.25	-1.49	0.24
ASSESSOR 2*	Lexis	-0.58	0.25	-1.40	0.23
ASSESSOR 3	Task	-0.31	0.24	-0.32	0.21
ASSESSOR 3	Organization	-0.81	0.26	-0.41	0.21
ASSESSOR 3	Grammar	-1.03	0.25	-0.47	0.22
ASSESSOR 3	Lexis	-0.77	0.25	-0.35	0.22
ASSESSOR 4	Task	0.08	0.24	0.30	0.21
ASSESSOR 4	Organization	-0.23	0.25	0.26	0.21
ASSESSOR 4	Grammar	-0.52	0.25	-0.08	0.22
ASSESSOR 4	Lexis	-0.52	0.25	-0.06	0.22
ASSESSOR 5	Task	0.47	0.24	0.30	0.21
ASSESSOR 5	Organization	0.29	0.25	0.08	0.21

ASSESSOR 5	Grammar	-0.08	0.25	0.01	0.22
ASSESSOR 5	Lexis	-0.07	0.25	0.27	0.22
ASSESSOR 6	Task	0.63	0.24	0.39	0.21
ASSESSOR 6	Organization	0.48	0.25	0.21	0.21
ASSESSOR 6*	Grammar	1.07	0.26	0.36	0.22
ASSESSOR 6	Lexis	0.51	0.25	0.32	0.22
ASSESSOR 7	Task	1.97	0.26	1.53	0.25
ASSESSOR 7	Organization	1.81	0.28	1.45	0.25
ASSESSOR 7	Grammar	1.75	0.27	1.30	0.25
ASSESSOR 7	Lexis	1.16	0.26	1.27	0.24
*Rasch measures differ by greater than 2x standard error					

Table 6. Average Rasch values for each assessor for each component

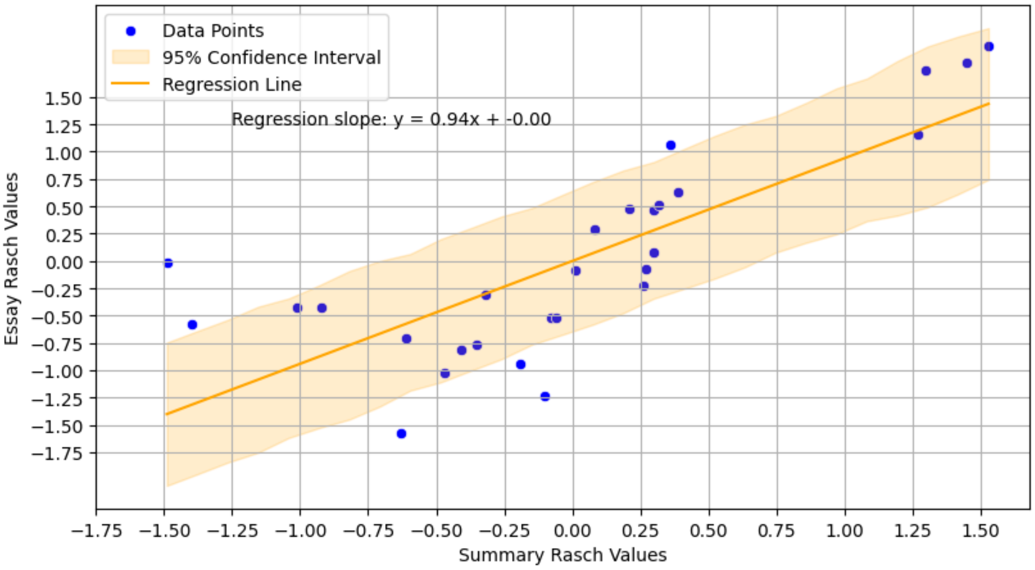


Figure 6. Average Rasch values for each component for each assessor with 95 percent CI

The regression slope in Figure 6 indicates a nrarly 1:1 relationship between Summary and Essay Rasch values. The relationship shows a strong positive correlation (.78, p < .01), showing that as Summary Rasch values increase, Essay Rasch values increase by a similar amount. Assessors' Rasch measures are very consistent across the two writing tasks, meaning their relative leniency/severity remains similar whether rating Summaries or Essays. Note that the assessor facets were centered at zero for both analyses, resulting in the regression line intersecting 0,0. The shaded area in Figure 7 represents the 95 percent confidence interval of the regression slope. This was calculated by multiplying the average standard error by two. There are six data points outside the shaded area, meaning that assessor behavior for that component is different in the Essay and the Summary by more than two standard errors. These six data points are marked in Table 7. Assessor 1 accounts for three out of the six data points for task fulfilment, organization and lexis. However, for the remaining six assessors, there is no systematic difference in how the different components are used across the two tasks.

Finally, correlations between the components of the rating scale were calculated based on the raw scores to examine the interrelationships between the components. The output is presented in Table 7.

	Task fulfilment	*Organization*	*Grammar*	*Lexis*
Task fulfilment	1			
Organization	0.92	1		
Grammar	0.89	0.90	1	
Lexis	0.90	0.89	0.94	1

Table 7. Correlation matrix for the four components of the rating scale for the Summary task

The data shows that all the criteria for the Summary task are strongly correlated at approximately .90. The correlation between grammar and lexis is higher, at .94, indicating that assessors perceive these two criteria as more closely related than they are to organization or task fulfilment.

Conclusions

This study represents the first step in providing validity evidence for an intertextual, reading-into-writing summary task to be used in the Oxford Test of English Advanced. We recognize that validation is an ongoing endeavor, and that further evidence will need to be presented to ensure that the task is fit for purpose in a test designed to be used for professional purposes and entrance to higher education. Evidence presented in this study shows that an intertextual reading-into-writing task is a good approach for an English test as the task was able to elicit responses from test takers at different levels of ability and presents evidence that assessors are able to score responses at similar levels of reliability to a more traditional Essay task.

The analytic rating scale incorporated reading-related and source use within the four individual components of task fulfilment, organization, grammar and lexis. The data analysis presented here provides additional support for the use of analytic rating scales to score Summary task performances and for the incorporation of source use within language use criteria, rather than creating a separate criterion specifically for source use. This avoids the problem identified by Lestari and Brunfaut (2023), who found that an independent 'reading for writing' discriminated less than other criteria.

Despite the integration of source use within existing criteria, scores for task fulfilment and organization were generally lower than scores for grammar and lexis in the Summary task. This may be due to test takers' proficiency level being below the difficulty of the task (Cumming, 2014) or test-takers' general lack of familiarity with this kind of intertextual reading-into-writing task (Chan, Inoue & Taylor, 2015). However, given that the test takers were specifically selected for participation in this study based on teachers' expert judgement of their level and subsequent recommendation, we consider the latter explanation to be the most plausible. As a result, test developers seeking to implement these kinds of tasks within high- stakes assessments must ensure that they are well-supported by supplementary materials, publicly available information about the construct and test specification, practice test tasks, hints, guides and advice on test-taking strategies to successfully complete the task.

A lack of familiarity with the task means it is likely to have a significant washback effect on test takers, who will adapt to the requirements of the task by regularly engaging in summarization practice. This is a hypothesized benefit of adopting this task, as Marzec-Stawiarska (2016) found that students who regularly summarized texts showed significantly greater improvement in reading comprehension compared to students who did more traditional reading activities like multiple choice questions. She also found that summary writing had a much more positive effect on developing reading skills for weaker readers.

An additional challenge identified in this study is the very high intercorrelations among the different components of the rating scale. This may indicate that assessors are struggling to distinguish between the components. Brown (2006) and Chan, Inoue & Taylor (2015) have previously noted this problem as endemic to analytic scales and identified this as an issue common to all criteria of this type. This is an issue which must be subject to ongoing monitoring in live testing, and asl speaks to the need for substantial assessor training to ensure that assessors pay attention to the multiple elements of the task.

The present research is not without limitations. The research presented here has been entirely quantitative. As part of the research, we have collected substantial feedback from both assessors and test takers to capture their perceptions of the suitability of the task for the stated purpose and how assessors felt using the criteria to score responses. However, for reasons of brevity, there is not sufficient space to present the findings from the qualitative aspects of the research here. Additionally, there is a need to explore the reading practices of test takers in more detail. A further avenue for research could also be to explore the reading phase in more detail in which test takers map intertextual relations before writing, which would make the synthesis process visible. Finally, an exciting avenue of exploration for tasks of this kind are natural language processing (NLP) approaches to automatically score source use elements like semantic overlap. Given that the task design already controls text length (up to 100 words), the task is well-designed for computational linguistic approaches which could complement human rating by identifying source integration behaviors automatically.

Appendices

Appendix A

CEFR	Score	Task fulfilment	Organization	Grammar	Lexis
C2	**7**	• response skilfully redrafts the main ideas with appropriate supporting details from both texts • response is consistently clear, sophisticated and appears effortless, with no redundancy • register is consistently appropriate for task purpose	• reconstructs ideas to produce a response with a natural flow • consistently coherent; well-structured with logical sequencing of ideas • uses sophisticated cohesive features appropriately at all times	• exploits grammatical resources creatively to write with a distinct voice • maintains consistent grammatical control to produce a very concise response • maintains a high level of accuracy throughout; errors are rare and only concern complex forms	• exploits lexical resources creatively with a high degree of sophistication • maintains consistent control of lexis and phrases to produce a very concise response • maintains a high level of accuracy of both lexis and phrases; errors are rare and difficult to spot
C1.2	**6**	Comfortably meets the positive descriptors of 5 and negative descriptors minimally if at all.			
C1.1	**5**	• response synthesises the main ideas with appropriate supporting details from both texts • response is clearly communicated with little redundancy • register is nearly always appropriate for task purpose	• reorganizes ideas in a logically connected way • consistently coherent; well-organized progression of ideas • uses appropriate cohesive features with rare instances of misuse	• exploits grammatical resources to adapt grammatical structures • maintains grammatical control to produce a concise response • maintains a good level of accuracy; occasional errors when adapting grammatical structures	• exploits lexical resources to adapt lexis • maintains lexical control to produce a concise response • maintains a good level of accuracy; occasional errors when adapting lexis
B2.2	**4**	Comfortably meets the positive descriptors of 3 and negative descriptors minimally if at all.			
B2.1	**3**	• response synthesises at least two main ideas with some supporting details from both texts • response is generally clearly communicated; shows awareness of task purpose • register is generally appropriate; response shows awareness of task purpose	• integrates ideas coherently while maintaining original sequence • generally coherent; able to connect ideas across sentences • uses simple cohesive features to link sentences, generally appropriately	• uses grammatical resources to paraphrase some grammatical structures • moderate grammatical control; response may lack conciseness • generally accurate; grammatical errors occasionally impede communication	• uses lexical resources to paraphrase some words and phrases • moderate lexical control; response may lack conciseness • generally accurate; lexical errors occasionally impede communication
B1.2	**2**	Comfortably meets the positive descriptors of 1 and negative descriptors minimally if at all.			
B1.1	**1**	• response includes at least one main idea from one text • response is not always clearly communicated • register is not always appropriate; limited awareness of task purpose	• reproduces ideas in their original sequence; little attempt to manipulate order of ideas • not always coherent; presents ideas as list of separate points • occasionally uses some simple cohesive features to link sentences	• uses grammatical resources to paraphrase in a simple fashion • limited grammatical control; response relies on original structures • sometimes inaccurate; errors may occur when paraphrasing simple structures	• uses lexical resources to paraphrase some words in a simple fashion • limited lexical control; response relies on original wording • sometimes inaccurate; errors may occur when paraphrasing frequent lexis
Below B1	**0**	Response does not fulfil all the positive descriptors of 1 (B1) **OR** task not attempted **OR** response is irrelevant (i.e. off-topic)			

Score cap: Use of input texts

- Only one source text is used: Band 2 (Task fulfilment and organization)

Score cap: Maximum word count

The word limit for this task is 100. Therefore:

- Up to 105 words: any score may be awarded.
- 105-120 words: Band 4 (all criteria)
- 121 words or more: Band 2 (all criteria)

Appendix B. Assessor measurement report

Assessor	T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtMea	PtMeExp	Exact Obs %	Agree Exp %
1	939	192	4.89	4.86	1	0.08	1.44	3.87	1.61	5.12	0.67	0.77	20.5	22.3
2	824	192	4.29	4.14	0.33	0.08	0.95	-0.48	0.95	-0.47	0.74	0.78	25.8	26.2
3	848	192	4.42	4.29	0.47	0.08	0.9	-1.04	0.9	-1	0.86	0.78	28.4	26.8
4	777	192	4.05	3.86	0.06	0.08	0.94	-0.62	0.94	-0.63	0.77	0.78	28	27.3
5	743	192	3.87	3.66	-0.14	0.08	1.11	1.1	1.11	1.09	0.79	0.78	27.1	27.2
6	697	192	3.63	3.39	-0.41	0.08	0.95	-0.46	0.95	-0.51	0.8	0.77	28.4	26.3
7	555	192	2.89	2.63	-1.3	0.08	0.7	-3.18	0.72	-2.98	0.77	0.74	20.3	19.4

Appendix C. Test taker measurement report

Test taker	T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtMea	PtMeExp	Discrim
30040	157	56	2.8	2.74	-1.23	0.15	1	0.05	1	0.09	0.65	0.58	0.93
30042	106	28	3.79	3.34	-0.54	0.19	0.45	-2.67	0.44	-2.72	0.69	0.57	1.61
60021	256	56	4.57	4.59	0.67	0.13	1.08	0.49	1.06	0.37	0.38	0.64	0.93
60022	107	28	3.82	4.24	0.35	0.19	0.64	-1.54	0.64	-1.55	0.76	0.57	1.44
81264	96	28	3.43	3	-0.91	0.2	0.38	-3.06	0.39	-2.97	0.63	0.55	1.64
81265	86	28	3.07	3.41	-0.46	0.2	0.45	-2.5	0.42	-2.72	0.79	0.54	1.61
81266	50	28	1.79	1.96	-2.4	0.27	0.96	-0.05	0.98	0.01	0.19	0.42	0.95
120033	148	56	2.64	2.58	-1.44	0.16	1.41	1.92	1.32	1.6	0.68	0.57	0.67
120035	136	28	4.86	4.43	0.53	0.19	0.65	-1.51	0.65	-1.5	0.83	0.58	1.37
120831	298	56	5.32	5.39	1.47	0.14	0.74	-1.5	0.8	-1.12	0.62	0.62	1.27
120832	110	28	3.93	3.47	-0.39	0.19	1.18	0.74	1.24	0.97	0.83	0.57	0.66
120854	68	28	2.43	2.69	-1.29	0.23	0.55	-1.84	0.58	-1.74	0.73	0.49	1.45
120855	69	28	2.46	2.14	-2.09	0.23	0.64	-1.42	0.72	-1.06	0.21	0.5	1.28
121146	312	56	5.57	5.66	1.77	0.15	0.94	-0.28	0.97	-0.12	0.71	0.6	1.1
121148	294	56	5.25	5.32	1.39	0.14	0.89	-0.56	0.86	-0.73	0.65	0.62	1.2
130200	273	56	4.88	4.91	0.98	0.14	0.7	-1.81	0.68	-1.95	0.77	0.63	1.38
130201	227	56	4.05	4.03	0.15	0.13	1.3	1.56	1.38	1.96	0.21	0.63	0.62
130203	254	56	4.54	4.55	0.63	0.13	1.74	3.46	1.84	3.85	0.43	0.64	0.15
250032	106	28	3.79	4.2	0.31	0.19	0.65	-1.48	0.66	-1.45	0.52	0.57	1.36
250034	329	56	5.88	5.97	2.18	0.16	1.59	2.63	1.99	4.03	0.51	0.58	0.49
250036	228	56	4.07	4.05	0.17	0.13	0.83	-0.94	0.81	-1.07	0.58	0.63	1.18
270018	82	28	2.93	2.55	-1.48	0.21	1.03	0.2	1.02	0.17	0.38	0.53	0.99
270020	157	56	2.8	2.74	-1.23	0.15	0.97	-0.09	0.92	-0.35	0.66	0.58	1.04
270021	128	28	4.57	5.02	1.09	0.19	1.55	1.94	1.53	1.89	0.56	0.58	0.4
270030	137	56	2.45	2.38	-1.72	0.16	1.18	0.95	1.16	0.88	0.53	0.55	0.79
270031	320	56	5.71	5.81	1.95	0.15	0.62	-2.22	0.63	-2.18	0.74	0.59	1.43
320025	229	56	4.09	4.07	0.19	0.13	1.23	1.23	1.2	1.1	0.65	0.63	0.85
320054	100	28	3.57	3.14	-0.76	0.19	0.39	-3.01	0.38	-3.09	0.82	0.56	1.69
320055	236	56	4.21	4.2	0.31	0.13	1.64	3.05	1.72	3.37	0.31	0.63	0.2
760038	168	56	3	2.94	-0.99	0.15	1.02	0.16	1.04	0.27	0.56	0.59	0.96
760039	116	28	4.14	4.58	0.67	0.19	0.75	-1.02	0.73	-1.1	0.81	0.58	1.37

Appendix D. Task measurement report

Task	T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtMea	PtMeExp	Discrim
Essay	2888	672	4.3	4.25	-0.43	0.04	0.9	-1.89	0.9	-1.98	0.81	0.8	1.12
Summary	2495	672	3.71	3.38	0.43	0.04	1.1	1.86	1.15	2.74	0.78	0.79	0.89

Appendix E. Rating scale measurement report (Summary task)

Component	T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ	PtMea	PtMeExp	Discrim
Task fulfilment	598	168	3.56	3.3	0.17	0.08	0.97	-0.23	1.04	0.41	0.8	0.8	1.02
Organization	609	168	3.63	3.36	0.06	0.08	0.98	-0.11	1.05	0.46	0.8	0.8	1
Grammar	647	168	3.85	3.71	-0.14	0.08	0.96	-0.31	1.01	0.16	0.81	0.81	1.04
Lexis	641	168	3.82	3.68	-0.09	0.08	0.95	-0.44	1.02	0.21	0.81	0.81	1.04

References

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573. https://doi.org/10.1007/BF02293814

Barkaoui, K. (2011). Think-aloud protocols in research on Essay rating: An empirical study of their veridicality and reactivity. Language Testing, 28(1), 51–75. https://doi.org/10.1177/0265532210376379

Brown, A. A. (2006). An examination of the rating process in the revised IELTS Speaking Test. https://https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report2.ashx

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. Assessing Writing, 26, 20–37. https://doi.org/10.1016/j.asw.2015.07.004

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume. Council of Europe. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Crossley, S. A., Wan, Q., Allen, L., & McNamara, D. S. (2023). Source inclusion in synthesis writing: An NLP approach to understanding argumentation, sourcing, and Essay quality. Reading and Writing, 36, 1053-1083. https://doi.org/10.1007/s11145-021-10221-x

Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), The companion to language assessment (pp. 216–229). Wiley-Blackwell. https://doi.org/10.1002/9781118411360

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. Assessing Writing, 21, 56–73. https://doi.org/10.1016/j.asw.2014.03.002

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In: L. Hamp-Lyons (Ed.), Assessing Second Language Writing in Academic Contexts (pp. 241–276). Norwood, NJ: Ablex.

Hyland, K. (2009). Academic discourse: English in a global context. Continuum.

Intersegmental Committee of the Academic Senates of the California Community Colleges and the University of California (2002) Academic Literacy: A Statement of Competencies Expected of Students Entering California's Public Colleges and Universities. Available at https://icas-ca.org/

Knoch, U., Zhang, B. Y., Elder, C., Flynn, E., Huisman, A., Woodward-Kron, R., Manias, E., & McNamara, T. (2020). ‘I will go to my grave fighting for grammar’: Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. Assessing Writing, 46, 100488. https://doi.org/10.1016/j.asw.2020.100488

Leijten, M., Van Waes, L., Schrijver, I., Bernolet, S., & Vangehuchten, L. (2019). Mapping master's students' use of external sources in source-based writing in L1 and L2.

Lestari, S. B., & Brunfaut, T. (2023). Operationalizing the reading-into-writing construct in analytic rating scales: Effects of different approaches on rating. Language Testing, 40(3), 684-722. https://doi.org/10.1177/02655322231155561

Li, J. (2014). Examining genre effects on test takers’ summary writing performance. Assessing Writing, 22, 75-90. https://doi.org/10.1016/j.asw.2014.08.003

Linacre, J. M. (2023a). Facets (Version 3.71.4) [Computer software]. https://www.winsteps.com/facets.htm

Linacre, J. M. (2023b). Facets program manual 3.86. https://www.winsteps.com/a/Facets-Manual.pdf

Linacre, J. M. (2023c). Winsteps tutorial 3. https://www.winsteps.com/a/winsteps-tutorial-3.pdf

Linacre, J. M. (2023d). Winsteps [Computer software]. https://www.winsteps.com/winsteps.htm

Lunz, M. E., Wright, B. D. and Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 3, 331-345.

Marzec-Stawiarska, M. (2016). The influence of summary writing on the development of reading skills in a foreign language. System, 59, 90-99. https://doi.org/10.1016/j.system.2016.04.006

McNamara, T. (1996). Measuring Second Language Performance. London and New York: Longman.

McNamara, T., Knoch, U., & Fan, J. (2019). Fairness, justice, and language assessment. Oxford University Press.

Nelson, N., & King, J. R. (2022). Discourse synthesis: Textual transformations in writing from sources. Reading and Writing, 35, 769–808. https://doi.org/10.1007/s11145-021-10243-5

Owen, N. (2016). An evidence-centered approach to reverse engineering: Comparative analysis of IELTS and TOEFL iBT reading sections, unpublished PhD thesis, University of Leicester.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. Assessing Writing, 17(1), 18–34. https://doi.org/10.1016/j.asw.2011.09.002

Sawaki, Y.; Stricker, L. J.; Oranje, A. H. (2009). Factor structure of the TOEFL Internet based test. Language Testing, 26 (1), 5 30. https://doi.org/10.1177/0265532208097335

Shaw, S. D., & Weir, C. J. (2007). Examining writing: Research and practice in assessing second language writing. Cambridge University Press.

Spivey, N. N. (1988). Discourse synthesis: Constructing texts in reading and writing. University of Delaware Press.

Studies in Second Language Acquisition, 41(3), 555-582. https://doi.org/10.1017/S0272263119000251

Weigle, S. C., Yang, W., & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. Language Assessment Quarterly, 10(1), 28-48. https://doi.org/10.1080/15434303.2011.642040

Weir, C.J., & Chan, S. (2019). Trends in language assessment research and practice: The view from Language Testing 1986-2016. Language Testing, 36(3), 349-363. https://doi.org/10.1177/0265532219831477

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. MESA Press.

Yu, G. (2013). From integrative to integrated language assessment: Are we there yet? Language Assessment Quarterly, 10(1), 110–114. https://doi.org/10.1080/15434303.2013.766744