CEFR JOURNAL—RESEARCH AND PRACTICE
VOLUME 8

Designing and validating an intertextual reading-into-writing summary task: A CEFR-aligned approach using the 2022 Handbook

Nathaniel Owen, Oxford University Press, Great Britain
Oliver Bigland, Oxford University Press, Great Britain

https://doi.org/10.37546/JALTSIG.CEFR8-3

This article reports on the design, development and CEFR-alignment of an innovative intertextual reading-into-writing summary task for the Oxford Test of English Advanced, targeting CEFR levels B2-C1. The summary task requires test takers to read two texts on the same topic (300 words total) and synthesize information into a 100-word summary. The study utilized the 2022 Aligning language education with the CEFR: A handbook (Handbook) to inform methodological decision-making throughout the development and alignment process, adopting an examinee-centred approach to validation. Data from a CEFR alignment panel (n = 7) and a larger-scale pilot study (n = 215) were analysed using many-facet Rasch measurement to investigate task performance, assessor reliability, and concurrent validity with a traditional essay task. Results indicate strong reliability (ICC = .87), equivalent to a traditional essay writing task, and demonstrate the task's effectiveness in distinguishing B2 from C1-level performances. The study provides evidence for the utility of the 2022 Handbook in guiding CEFR alignment methodology and provides evidence that summary writing tasks are a viable alternative to traditional essay writing assessments in high-stakes contexts.

Keywords: CEFR alignment, summary writing, integrated skills, reading-into-writing, mediation, Handbook 2022

1 Introduction

Language testing has historically been predicated on a "four skills" approach, treating listening, speaking, reading, writing as isolated competencies (Lado 1961). However, this traditional approach fails to reflect the integrated nature of language use in real-world contexts, particularly in academic and professional settings where skills are routinely combined to complete specific tasks (Gebril and Plakans 2014; Plakans and Gebril 2012; Yu 2013). The *Common European Framework of Reference for Languages* (CEFR) (Council of Europe [CoE] 2001) advocates for "modes of communication" rather than individual "skills" and has evolved to encompass concepts such as "interaction" and "mediation" within the *CEFR Companion Volume* (CEFR/CV; CoE 2020) which emphasize combining skills for describing overall language proficiency.

Against this backdrop, Oxford University Press has developed an intertextual reading-into-writing summary task for the *Oxford Test of English Advanced*, designed to measure CEFR levels B2-C1, using the *Companion Volume* to inform task design. This task requires test takers to read two texts, select, paraphrase and synthesize information into a single piece of writing, closely resembling the kinds of language use tasks examinees encounter in academic and professional contexts (Sawaki et al. 2009).

As this task was being developed, the publication of the 2022 Handbook *Aligning Language Education*

with the CEFR (British Council, UKALTA, EALTA and ALTE 2022, hereafter referred to as the Handbook) marked a significant development in CEFR alignment methodology, providing updated guidance for aligning language tests, curricula, and course content to the CEFR. The Handbook emphasizes the importance of the Comprehensive Learning System (CLS) approach, which advocates for close alignment of curriculum, delivery, and assessment elements (O'Sullivan 2020).

This article reports on the development and validation of this summary task, demonstrating how the 2022 Handbook informed methodological decisions throughout the alignment process. Specifically, the Handbook distinguishes between examinee-centred and test-centred approaches to validation. The former approach requires the collection of test taker performance samples and having them evaluated by independent, external participants. The latter approach focuses more specifically on the evaluation of test design and content (Handbook: 49). This study provides evidence for the utility of the Handbook's recommendations in guiding CEFR alignment for mediation-related test tasks.

2 Literature review and theoretical framework

2.1 Mediation in the CEFR and the Handbook

The concept of mediation, central to the CEFR Companion Volume, refers to "a social and cultural process of creating conditions for communication and cooperation" (Council of Europe 2020: 106) and includes both cross-linguistic mediation and mediation within a target language, chiefly concerned with facilitating the communicative needs of others. Mediation often occurs across modalities, where written output may involve processing and relaying the message of a spoken text or synthesizing multiple sources. The original CEFR framework (CoE 2001) established a theoretical foundation for mediation in language assessment, while the later CEFR/CV (CoE 2020) expanded the concept to better reflect contemporary language use in electronic and multilingual contexts. The 2020 CEFR/CV and 2022 Handbook also emphasize mediation as a crucial component of CEFR alignment, particularly for integrated-skills tasks. Figueras et al. (2022) provide a comprehensive overview of the Handbook's development and its significance for CEFR alignment methodology.

2.2 The intertextual reading-into-writing construct

The purpose of developing an integrated reading-into-writing task is to better represent the kinds of activities that are fundamental to academic and professional language proficiency domains. When reading for writing, language users adopt appropriate reading strategies to construct models of text structure, construct textual and intertextual representations that allow them to select, evaluate, and use information according to the writing purpose (Weigle et al. 2013). This process represents a form of discourse synthesis (Nelson and King 2022), which refers to operations such as organizing, selecting, and connecting content from multiple sources on the same topic. For summary writing tasks, evaluation criteria should examine content transformation and degree of source in addition to traditional criteria content such as organization, grammar or vocabulary. Higher scores should be awarded for making explicit links across sources, especially where such links may only be implied in the original source texts.

2.3 Designing intertextual reading-into-writing summary tasks

The integrated nature of a summary task means that test developers are required to address multiple design considerations, including input (what test takers are required to read), output (what test takers are required to do), and how responses will be scored. Regarding input, Li (2014) found 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. Students performed better when summarizing expository texts compared to narrative texts, as expository texts contain more

explicit topic sentences and hierarchical structures compared to narrative texts. In traditional writing tasks such as essays, text length is often the strongest predictor of test taker performance (Crossley et al. 2023). However, summary tasks require test takers to select information, meaning lengthier responses may show less evidence of test takers' ability to discriminate between main and supporting information. Summary tasks may therefore benefit from an *upper* word count rather than a minimum word count to ensure idea selection and synthesis rather than text reproduction and paraphrasing.

Regarding scoring, analytic rating scales are generally preferred over holistic scales due to the complexity of cognitive processing involved in task completion. Developers must decide whether source use is a separate scale or integrated into descriptors for other rating scale components. Lestari and Brunfaut (2023) compared the use of a scale with a separate criterion for reading/source use with a scale which integrated source use with writing criteria, finding that both scales functioned well, but the separate criterion offered greater transparency to raters.

To date, existing tests of English used for university admission or professional purposes have largely 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. The brief discussion here demonstrates that developing such a task represents a significant challenge, complicated by the requirement of developing carefully controlled input, identifying explicit task requirements (i.e., what test takers are required to do with the input), and how to offer support to assessors who must navigate between task requirements, input texts and the test taker's response. This complexity likely explains the dearth of such tasks. However, this omission has resulted in the proliferation of English language tests which inadequately reflect real-world language use in academic and professional domains.

2.4 The 2022 Handbook's contribution to CEFR alignment

The 2022 Handbook provides updated guidance on aligning language tests to the CEFR, building on *Relating Language Examinations to the Common European Framework of Reference for Languages: A Manual* (Council of Europe 2009, hereafter referred to as the Manual) but incorporating developments from the 2020 CEFR Companion Volume. The Handbook emphasizes the importance of the examinee-centred approach for integrated skills assessment, which involves collecting test taker performance samples and scoring them using established systems by external participants. The Handbook's five-stage alignment process (familiarization, specification, standardization, standard setting, and validation) provides a structured approach to ensuring CEFR alignment, consistent with that found in the Manual. The examinee-centred approach is particularly relevant for integrated skills tasks, as it allows for the collection of authentic performance data to investigate the complex task completion processing involved.

3 Methodology

3.1 Research design and research questions

This study employed a mixed-methods approach combining quantitative analysis of test performance data with qualitative assessment of task design and validation procedures. The research design followed the five-stage alignment process outlined in the 2022 Handbook: familiarization, specification, standardization, standard setting, and validation. The data reported in this study represent the standard setting and validation phases of alignment.

The study addressed the following research questions:

RQ1: To what extent does the summary task demonstrate equivalent reliability to traditional essay tasks in measuring writing proficiency at CEFR levels B2-C1?

RQ2: How effectively do assessors score summary task responses using an analytic rating scale compared to traditional essay tasks?

RQ3: What is the concurrent validity between summary task performance and traditional essay task performance?

RQ4: To what extent does the summary task discriminate between B2 and C1 level performances according to CEFR standards?

RQ5: How does the difficulty level of the summary task compare to traditional essay tasks, and what factors contribute to any differences?

Data from the online CEFR alignment validation panel (n = 7) and subsequent large-scale pilot study (n = 215) were used to address the research questions. RQ1 and RQ2 were addressed using pilot study data, RQ3 was addressed using data from test takers who completed both a summary and an essay task in the pilot study, and RQ4 and RQ5 were addressed using CEFR alignment validation data from the panel.

3.2 Task design

The summary task was designed using CEFR mediation descriptors at B2 and C1 in addition to detailed domain analysis and the recommendations of the 2022 Handbook. Test takers are presented with two texts on the same topic (300 words total) and required to synthesize information into a summary. The task parameters and features include:

- Two source texts of different genres, one textbook extract and one lecture transcript (approximately 150 words each)
- Clear instructions emphasizing synthesis rather than reproduction
- A 100-word limit to ensure idea selection and transformation
- A glossary of low-frequency lexis to support comprehension
- 20-minute time limit including both reading and writing time

For more detail regarding task design and scoring, please see the online test specifications (Oxford University Press 2025).

3.3 Participants

A total of 665 test takers participated in the pilot study. A total of eight assessors marked the Speaking and Writing test scripts. 314 of the test takers received scores from a minimum of two assessors. However, of these 314, only 215 received marks for both essay and summary writing tasks from a minimum of two assessors. As a result, assessor data presented in the findings (RQ2) is the output of analysis of the 314 test takers. Concurrent validity and reliability research questions (RQ1, RQ3 and RQ5) are addressed using data from the 215 test takers for direct comparability. Piloting took place in Spring 2024. The pilot study collected data across multiple test administrations in Spring 2024, with detailed findings reported in the Oxford University Press pilot study report (Owen 2024b).

As part of ethical approval, participants were not required to complete biodata entries as a condition of participation. As a result, biodata collection was partial. Of the 215 test takers reported for reliability and concurrent validity research questions, 185 responded to the biodata questions. The first language reported by the cohort is dominated by Turkish (86) and Spanish (Castilian) (57), followed by Italian (16), Arabic (9), German (4), Portuguese (3) and one each for Catalan-Valencian, Portuguese, Czech, Kurdish, Luxembourgish, Dutch-Flemish and English (one teacher participant). Gender distribution shows 104 females, 78 males, and 3 “prefer not to say”. Ages range from 7 to 66 years (mean ≈ 22.8), with most participants between 16 and 25 years. Test takers were recruited through test centres and selected based on expert judgment of their B2–C1 level proficiency carried out by teachers at their respective test centres.

For the CEFR alignment validation panel, seven assessors participated in the study. The panel consisted of five male and two female assessors, all English L1 speakers with extensive experience in language assessment. Three assessors held PhDs in language testing, while the remaining four possessed extensive experience in English language teaching, item writing, and materials development. All standard setting activities were undertaken by assessors independently and all assessors received the same standard setting materials. Samples were selected from pretesting, which had occurred in summer 2022. Panel members judged 48 samples of writing (24 essay and 24 summary responses) across four pretests, with each pretest containing an essay and a summary task (six responses per pretest). The CEFR alignment panel and pilot study used the same analytic rating scale with four components: Task fulfillment, Organization, Grammar, and Lexis. Each response received four scores (maximum score = 28). Standard setting data is used to address RQ4. The CEFR alignment validation procedures and results are documented in detail in the CEFR alignment report for *Oxford Test of English Advanced* Writing and Speaking modules (Owen 2024a).

3.4 Data analysis

Data for both the pilot study and the CEFR alignment were analysed using many-facet Rasch measurement (MFRM) using a Rasch-Masters partial credit model within the program FACETS v3.84 (Linacre 2023). A five-facet model was adopted (assessors, test takers, task, pretest, component) with an eight-point rating scale (0-7). Reliability was assessed using the intraclass correlation coefficients, rater agreement, separation and strata output from the FACETS analysis.

4 Findings

4.1 RQ1: Task performance and reliability

The summary task demonstrated strong reliability across both pilot study and CEFR alignment validation. The summary task achieved a reliability value of .87, exactly equal to the traditional essay task (.87) (n = 215), indicating that the integrated nature of the task does not compromise measurement reliability, supporting the viability of intertextual reading-into-writing tasks in high-stakes assessment contexts. Inter-rater agreement was 35.1% for exact agreements, which is consistent with expected levels for subjective assessments and identical to the 35.1% agreement achieved for essay tasks. The separation and strata indices (8.41 and 11.54 respectively) for the summary task indicated strong proficiency differentiation by assessors while maintaining consistency in rating standards. These values are comparable to those achieved for essay tasks (12.40 and 16.87), suggesting that assessors applied rating criteria to integrated skills tasks with similar consistency to traditional essay writing tasks.

4.2 RQ2: Assessor performance

Many-facet Rasch analysis revealed strong assessor performance for the summary task. Table 1 presents the assessor performance statistics for Writing Script 2 (Summary task) from the pilot study data.

Table 1. *Assessor performance statistics (N.B. data for n = 314 test takers)*

Assessor	T.Score	T.Count	Obs.Avge	FairMAvge	Measure	S.E.	InfitMS	OutfitMS	PtMea	Discrim
1	2800	1256	2.23	2.21	.92	.04	.89	.88	.81	1.11
2	2769	820	3.38	3.15	-.72	.04	1.36	1.36	.84	.61
3	2240	852	2.63	2.43	.48	.05	.94	.93	.83	1.06
4	1708	428	3.99	3.33	-.99	.06	1.09	1.25	.87	.71
5	1867	600	3.11	2.69	.02	.05	.91	1	.86	1.06
6	2829	1164	2.43	2.42	.51	.04	.90	.93	.84	1.07
7	147	68	2.16	2.51	.33	.17	.73	.71	.78	1.29
8	1809	528	3.43	3.04	-.55	.05	.81	.79	.86	1.23
Model, Populn: RMSE .08 Adj (True) S.D. .63 Separation 8.41 Strata 11.54 Reliability (not inter-rater) .99										
Model, Sample: RMSE .08 Adj (True) S.D. .68 Separation 8.99 Strata 12.33 Reliability (not inter-rater) .99										
Model, Fixed (all same) chi-squared: 1436.4 d.f.: 7 significance (probability): .00										
Model, Random (normal) chi-squared: 7.0 d.f.: 6 significance (probability): .33										
Inter-Rater agreement opportunities: 12920 Exact agreements: 4530 = 35.1% Expected: 4685.8 = 36.3%										

The analysis shows that assessors were able to score summary responses effectively using the analytic rating scale. Most assessors demonstrated good fit statistics, with Infit and Outfit mean square values close to the ideal range of 0.7-1.3. Assessor 2 showed slightly higher fit values (1.36 for both Infit and Outfit), indicating some inconsistency in rating patterns, while

Assessor 7 demonstrated excellent fit (0.73 and 0.71 respectively) despite having a smaller sample size. Point-measure correlations ranged from 0.78 to 0.87, indicating strong correlation between assessors' ratings and expected ratings. Discrimination values ranged from 0.61 to 1.29, with most assessors achieving values above 1.0, indicating effective differentiation between different levels of performance.

4.3 RQ3: Concurrent validity

To assess concurrent validity, we compared performance on the summary task with traditional essay tasks. Figure 1 shows the relationship between summary and essay scores for test takers who completed both tasks and received scores from more than one assessor. Total scores represented are summed from fair mean averages output from Rasch-Masters partial credit FACETS analysis of the analytically scored dataset.

The scatter plot shows the relationship between scores on Writing Script 1 (Essay) and Writing Script 2 (Summary) for n = 215 test takers in the pilot study who completed both tasks. The r-squared value indicates that approximately 64% of the variance in Essay scores can be explained by the variance in Summary scores. A correlation of .80 suggests that the summary task measures similar underlying writing ability to traditional essay tasks, while also capturing additional skills related to reading comprehension and information synthesis. Test takers generally received slightly lower scores for the summary task compared to essay tasks. This pattern is visible in the scatter plot, with the regression slope and most data points falling below the ideal x = y line (marked in red), indicating that test takers typically scored higher on essay tasks than summary tasks. This difference is attributed to the additional

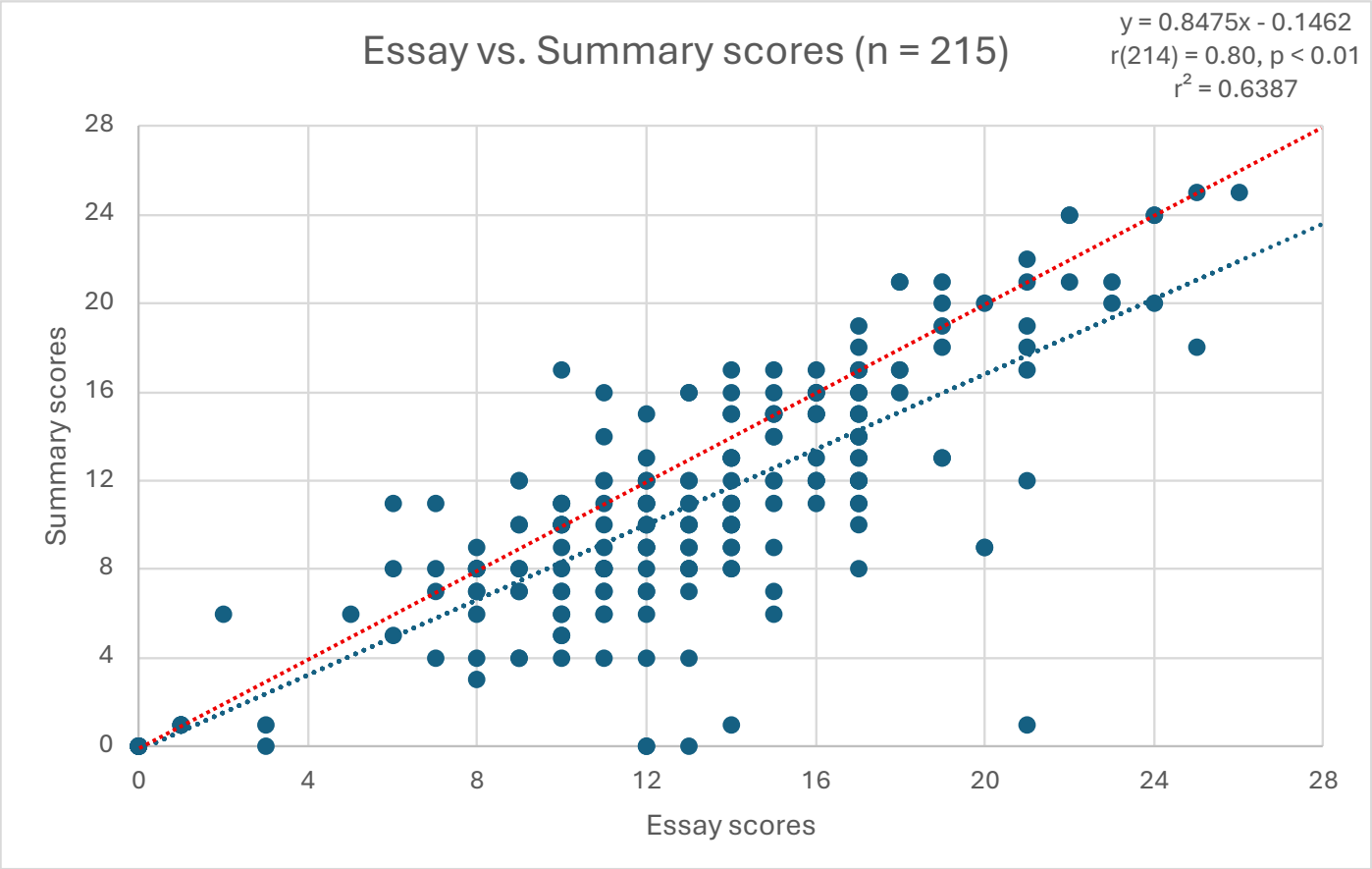

Figure 1. Essay vs. Summary scores for pilot participants (n = 215)

cognitive demands of processing multiple sources and synthesising information, as well as test takers' relative unfamiliarity with the integrated task type.

4.4 RQ4: CEFR level discrimination

Data from the CEFR alignment panel activities showed that the summary task effectively discriminated between B2 and C1 level performances. Forty-eight sample performances (24 essay and 24 summary) which had been marked internally were used in the CEFR standard setting activity. Eighteen out of 24 summary samples were awarded the same CEFR band by panel assessors as awarded during internal pretesting, demonstrating strong alignment with CEFR standards. The correlation between internal and external assessor scores was .94 for the summary task, indicating very strong agreement between different rating panels. The overall level of agreement for awarding CEFR bands was .77. The CEFR alignment validation showed that the summary task can be used to effectively distinguish between levels B2 and C1. The findings suggest that the task successfully measures the intended CEFR levels and that assessors can consistently apply CEFR standards to summary task responses.

4.5 RQ5: Task difficulty analysis

Rasch analysis revealed that the summary task was approximately one logit more challenging than traditional essay tasks. The difficulty measures for the two tasks were 0.43 logits for the essay task and -0.43 logits for the summary task, representing a difference of approximately 0.86 logits. This difference is statistically significant ($\chi^2 = 212.3$, d.f. = 1, $p < .001$). The increased difficulty of the summary task is attributed to the additional cognitive demands of processing multiple sources and synthesizing information into a single test. Also, test takers' relative unfamiliarity with the integrated task type may

contribute to the increased difficulty, as they lack experience with this type of assessment format. The difficulty difference was consistent across different administrations and assessor panels, suggesting that it reflects a genuine difference in task complexity rather than measurement error. This finding supports the validity of the summary task as a measure of advanced language proficiency, as it successfully differentiates between different levels of ability within the B2-C1 range.

5 Discussion and conclusions

5.1 Implications for test design

The success of the summary task demonstrates that integrated skills assessment can achieve reliability levels equivalent to traditional essay-style tasks. The task's effectiveness in discriminating between B2 and C1 levels supports its use in CEFR-aligned assessment contexts. The additional cognitive demands of the task, reflected in its higher difficulty level, may provide better measurement of advanced language proficiency by requiring test takers to demonstrate higher-order processing skills. The success of the analytic rating scale with integrated source use criteria suggests that separate source use scales may not be necessary for integrated tasks. This finding aligns with the recommendations of Lestari and Ho (2023) and provides practical guidance for rating scale development.

5.2 Contribution to CEFR alignment methodology

This study demonstrates the utility of the 2022 Handbook in guiding CEFR alignment methodology, particularly in the context of online alignment activities. The Handbook's examinee-centred approach proved particularly valuable for validating integrated skills tasks, as it allowed for the collection of authentic performance data that reflects the complex nature of real-world language use.

The five-stage alignment process outlined in the Handbook (15-17) provided a structured approach to ensuring CEFR alignment that was well-suited to fully online implementation. The Handbook's concise format, compared to the more extensive 2009 Manual, made the alignment process more navigable and manageable for participants. The Handbook's excellent organization and cross-referencing with the Companion Volume and Manual (8-9) proved particularly valuable during the alignment process. When participants required additional information about specific stages or procedures, they could easily locate relevant sections in the supporting documents, ensuring that the alignment process remained comprehensive despite the Handbook's more concise format.

5.3 Online alignment implementation

All CEFR alignment activities in this study were conducted online through Microsoft Teams (Microsoft Corporation n.d.) webinars, representing a significant departure from traditional face-to-face alignment procedures. This online approach offered several advantages that enhanced the alignment process. The Teams platform facilitated easy collection and distribution of materials through well-organized Teams folders, ensuring that all participants had consistent access to necessary documents and resources throughout the alignment process.

The online format also mitigated potential group dynamics issues that can occur in face-to-face settings. Individual participants working remotely were less likely to encounter pressure from more experienced team members and simply defer to their expertise, i.e., conformity bias or groupthink (Janis 1972). This reduced the risk of dominant personalities influencing group decisions and ensured that all participants could contribute equally to the alignment process. The asynchronous nature of some online activities allowed participants to work independently on rating tasks before coming together for discussion, further reducing the potential for group pressure to influence individual judgments.

5.4 Limitations of the Handbook and future directions

While noting that the layout of the Handbook lends itself well to online alignment processes, the 2022 Handbook has several limitations that became apparent during the online alignment process. The Handbook provides limited information about conducting online alignment procedures and how these will or should differ from face-to-face alignment panels, despite the increasing prevalence of remote work and virtual collaboration in language assessment. This gap in guidance may prove challenging organizations seeking to conduct future alignment activities in online environments. Future editions of the Handbook should seek to expand information around online alignment processes.

The Handbook also offers limited information about the impact of technology on alignment processes. New approaches such as AI or LLM-informed alignment and adaptive comparative judgment methods are emerging as potential alternatives to traditional alignment panels. The Handbook would benefit from addressing technological developments and their implications for future CEFR alignment methodology. Additionally, the Handbook provides limited technical support for those seeking to align to the CEFR. The establishment of the CEFR as an online API could significantly enhance alignment processes by providing standardized access to CEFR descriptors and facilitating automated alignment procedures. Such an API could enable real-time validation of alignment decisions and provide immediate access to relevant CEFR documentation, potentially streamlining the alignment process and reducing the time required for comprehensive alignment activities.

6 References

British Council, UKALTA, EALTA, & ALTE. 2022. *Aligning language education with the CEFR: A handbook*. In Neus Figueras, Barry O'Sullivan, Nick Saville, Lynda Taylor, & David Little (eds.). http://www.ealta.eu.org/documents/resources/CEFR%20alignment%20Handbook.pdf (accessed 20 Nov 2025).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. 2009. *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual*. Strasbourg: Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr (accessed 20 Nov 2025).

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion Volume*. Strasbourg: Council of Europe.

Crossley, Scott A., Qian Wan, Laura K. Allen & Danielle S. McNamara. 2023. Source inclusion in synthesis writing: An NLP approach to understanding argumentation, sourcing, and essay quality. *Reading and Writing* 36. 105-1083. https://doi.org/10.1007/s11145-021-10221-x.

Figueras, Neus, David Little & Barry O'Sullivan. 2022. Aligning language education with the CEFR: A handbook. *CEFR Journal, 5,* 1-10. https://doi.org/10.37546/JALTSIG.CEFR5-1

Gebril, Atta & Lia Plakans. 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.

Janis, Irving L. 1972. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascos.* Boston: Houghton Mifflin.

Lado, Robert. 1961. *Language testing: The construction and use of foreign language tests.* New York: McGraw-Hill.

Lestari, Santi Budi & Tineke Brunfaut. 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, Jiuliang. 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, John M. 2023. *Facets* (Version 3.71.4) [Computer software]. https://www.winsteps.com/facets.htm (accessed 20 Nov 2025).

Microsoft Corporation. n.d. *Microsoft Teams* [Computer software]. Retrieved from https://www.microsoft.com/en-us/microsoft-teams/group-chat-software (accessed 20 Nov 2025).

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

O'Sullivan, Barry. 2020. *The Comprehensive Learning System.* London: British Council.

Owen, Nathaniel. 2016. *An evidence-centred approach to reverse engineering: Comparative analysis of IELTS and TOEFL iBT reading sections*. University of Leicester PhD thesis.

Owen, Nathaniel. 2024a. *Oxford Test of English Advanced CEFR alignment report: Speaking and Writing.* https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/4-ref-0011-cefr-alignment-report-sandw-for-website.pdf (accessed 20 Nov 2025).

Owen, Nathaniel. 2024b. *Oxford Test of English Advanced pilot study report.* https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/4-ref-0030-pilot-study-report-2024-for-website.pdf (accessed 20 Nov 2025).

Oxford University Press. 2025. *Oxford Test of English Advanced test specifications.* Oxford: Oxford University Press. https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/oxford-test-of-english-advanced-test-specifications.pdf (accessed 20 Nov 2025).

Plakans, Lia & Atta Gebril. 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, Yasuyo, Lawrence J. Stricker & Andreas H. Oranje. 2009. Factor structure of the TOEFL Internet-based test. *Language Testing* 26(1). 5-30. https://doi.org/10.1177/0265532208097335.

Weigle, Sara Cushing, Weiwei Yang & Megan Montee. 2013. Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly, 10(1),* 28-48. https://doi.org/10.1080/15434303.2012.750659.

Weir, Cyril J. & Sathena Hiu Chong Chan. 2019. Trends in language assessment research and practice: The view from *Language Testing* 1986–2016. *Language Testing* 36(3). 349-363. https://doi.org/10.1177/0265532219826396.

Yu, Guoxing. 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.

7 Biographies

Nathaniel Owen is Senior Research and Analysis Manager at Oxford University Press. He holds a PhD in language testing from the University of Leicester specializing in L2 reading processes. His research interests and publications include the interface of language testing and technology, developing integrated-skills tasks, big data analytics, the use of language tests in English-medium instruction contexts, research methods and widening participation in higher education.

Oliver Bigland holds an MA in Applied Linguistics from the University of Birmingham. His research interests include the design and evaluation of integrated skills tasks, the role of functional language in speaking assessments, and the identification and mitigation of bias in language testing. He is also interested in the practical application of Rasch measurement theory and computational methodologies, particularly Python-based data analysis, in the context of language assessment.