



Assessing Multimodal Viewing-to-Write Constructs: Task Design, Performance, Processing, and Rating

Tineke Brunfaut & Judit Kormos

To cite this article: Tineke Brunfaut & Judit Kormos (15 Dec 2025): Assessing Multimodal Viewing-to-Write Constructs: Task Design, Performance, Processing, and Rating, Language Assessment Quarterly, DOI: [10.1080/15434303.2025.2596374](https://doi.org/10.1080/15434303.2025.2596374)

To link to this article: <https://doi.org/10.1080/15434303.2025.2596374>



© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 15 Dec 2025.



[Submit your article to this journal](#)





[View related articles](#)



[View Crossmark data](#)

Assessing Multimodal Viewing-to-Write Constructs: Task Design, Performance, Processing, and Rating

Tineke Brunfaut ^a and Judit Kormos ^{a,b}

^aLancaster University, Lancaster, UK; ^bUniversity of Ljubljana, Ljubljana, Slovenia

ABSTRACT

To reflect the present-day prevalence of multimodal language uses, which are underrepresented and underresearched in second language formal assessments, we developed and investigated two novel types of viewing-to-write tasks and accompanying rating scales. These multimodal integrated tasks require watching a visual-aural input and producing a written text based on it, targeting the language functions of describing and comparing-contrasting. We administered four task versions to 134 EFL learners, with a subgroup of 20 participating in post-task recall interviews. Descriptive statistics, correlation analyses, and mixed-effects modelling of test scores and qualitative analyses of the recall interviews showed that our viewing-to-write tasks and rating scales elicit, reflect, and evaluate forms of multimodal integrated language use and are practical for measuring such abilities of intermediate-advanced level learners.

INTRODUCTION

While the so-called four-skills approach has historically been crucial in transforming models of second language (L2) competence and innovating L2 teaching and assessment, human communication has never been limited to single-skill usage. Situations have always existed in which meaning is constructed and established through a variety of combinations of aural, gestural, spatial, visual and/or written channels (Kress, 2010). Thus, language is often used not just in a singular aural or written mode, but multimodally, involving several channels at the same time. Sometimes, the different modes require the application of two or more language skills such as in what are typically labelled “integrated language uses” – e.g., reading and listening to write, or listening-to-speak. In other cases, a more “diverse” range of modes are utilized, for example, visual and written channels in designing or comprehending brochures with pictures, drawings, text, and layout features.

The current substantial and rapid increases in the spread, accessibility and functionalities of technology are scaling up the extent and frequency of multimodal language uses (Mills & Unsworth, 2017). For example, e-mail communication may centre on graphs included in the text, written chat exchanges may comment on embedded pictures/videos, someone may watch a breaking news clip and relay the news to a family member entering the room, etc. This prevalence of multimodal language uses also means that effective communication in

CONTACT Tineke Brunfaut  t.brunfaut@lancaster.ac.uk  School of Social Sciences, Linguistics and English Language, Lancaster University, Lancaster LA14YL, UK

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

social, educational and professional contexts nowadays requires multimodal literacy (Perniss, 2018) – the ability to successfully employ communication practices which involve two or more modes of meaning (Mills, 2016).

These realities of human communication have meant that theories of language and meaning-making have been expanding beyond linguistic code and language purely as speech/text, towards multichannel events including written, oral and visual semiotic systems. Also, given the range of potential semiotic resources and possible combinations of different modes, it has been recognised that multimodality can take many shapes and forms (literally and figuratively). Similarly, even when the same modes are involved, their exact nature and intermodal balance as well as the context of use will determine the nature of the multimodality. Consequently, we can conceptualise multimodality as existing on a continuum of weaker to stronger multimodality. For example, a conventional L2 listening task involves an audio passage with written-text or picture-based items, thus involving two modes (aural-written or aural-visual). However, all task elements focus on listening comprehension and purposely aim to minimize any other modes; thus, a listening task is strictly speaking multimodal, but a very weak form of it and is usually labelled as monomodal. A daily life activity that might more typically be understood as multimodal and would thus be further along the continuum, is for example a doctor talking a patient through a diagnosis by means of a scan image and patient-oriented brochure while also responding to the patient's questions, taking into account the latter's body language. Other forms of multimodality, despite communicating meaning, are less typically regarded as falling within the realm of linguistics or language education, for example a dance enactment of a story, with prevalent gestural, spatial and visual modes, and only the occasional spoken or written expression.

Another significant development in scholarly understandings of language use and meaning-making is that multimodality is not a collection or sequence of individual, independent, distinct modes. Therefore, previous discussions of muddled constructs (e.g., Weir, 1990) do not fully reflect the nature of multimodality. According to more recent conceptualizations of multimodality, the various modes are fundamentally integrated and transformed (e.g., Nelson & King, 2022). Although the extent and way in which modes are integrated and interact with a person's semiotic systems in a particular context might vary, multimodal integratedness should be viewed as the construct itself.

LITERATURE REVIEW

Following the realisation and recognition of the multimodal nature of most human communication, language and literacy pedagogies in many contexts have now embraced multimodal approaches, including in L2 settings. The last decade has therefore also seen a vast increase in research on multimodality in L2 learning and teaching contexts – see, for example, various journal Special Issues on this topic (Crawford Camiciottoli & Campoy-Cubillo, 2018; Early et al., 2015; Peters & Muñoz, 2020; Yi et al., 2020). This has not yet been paralleled, however, in L2 assessment research or practice, especially in the context of formal, summative assessments such as L2 tests. To date, few language tests – whether used in classrooms or in large-scale standardised contexts – reflect stronger forms of multimodality as present in real-world domains, let alone the vast range of multimodal forms of human communication. In this sense, current L2 testing practices are limiting

construct representation and score relevance of formal, summative assessments, and thus risking negative consequences. As Shohamy (2022) argued, if L2 testing wants to remain relevant to 21st century communication needs, not only are insights needed into the nature of multimodal language processing and production, but also into how this can be formally assessed.

While special issues by Whithaus (2014) and Tan et al. (2020) were dedicated to multimodal assessment, these focused on general educational assessment. In L2 contexts, Lim et al.'s (2022) systematic review on multimodality in English language classrooms indicated a disconnect between multimodal pedagogies/curricula and prevailing product-oriented, language-dominant formal assessments, with negative washback implications. Additionally, when L2 multimodal assessment tasks are adopted, rating approaches tend to operationalise “traditional” linguistic performance aspects (Lim et al., 2022). Similar conclusions were drawn in Zhang et al.'s (2023) review on L2 digital multimodal composing, highlighting that formal assessment of this skill remains underexplored. The main exceptions are forms of L2 testing that focus on one particular language skill but include non-written/non-aural elements in the input (e.g. video-listening tasks, graph- or picture-prompt writing/speaking tests). Such inputs feature on several standardised proficiency tests (e.g. Aptis, IELTS, PTE, TOEFL) and have been the focus of ample research (e.g. Liu & Aryadoust, 2024; Yu et al., 2010). However, these can often be considered limited forms of multimodal L2 assessment, as the items or ratings tend to operationalise the language skill in a narrow, single-modal sense.

An initial set of empirical studies on broader conceptualisations of multimodal L2 assessment, published recently, nevertheless indicate growing interest, albeit mostly in the context of alternative or more informal forms of assessment. It can be observed, however, that most such studies are situated in higher education, English for Academic/Specific Purposes (EAP/ESP) programmes, concern continuous/formative/project assessments and involve pair or groupwork. For example, Beltrán-Palanquez (2024), Cheung (2023) and Hafner and Ho (2020) all focused on rubrics for assessing digital multimodal composing. Beltrán-Palanquez (2024) illustrates rubric development for assessing video game narrative tasks (groupwork) in a B2-level ESP course at a Spanish university. Video game narratives serve to communicate the story of a proposed game and elements that will shape the gaming experience (e.g. characters, settings, music), clarifying meaning and cohesion for players. Narratives, resulting from activities such as oral brainstorming, negotiation, and resource searches, are presented as lengthy, structured written text comprising e.g. strategically selected pictures. To establish assessment criteria relevant to real-life communication, Beltrán-Palanquez's rubric covered linguistic characteristics of the writing, expression of content and interpersonal meaning (e.g. reference to visuals), and use of multimodal resources and their coherent integration to enhance overall meaning (e.g. visual support, music, typography, multimodal coherence and intersemiotic relationships). Cheung (2023) developed and evaluated four rubrics for an EAP formative assessment of pair-work academic blog tasks (including visuals/videos/etc.) at a Hong Kong college. Two rubrics (one for students, one for teachers) evaluated the multimodal collaborative process and two focused on the output produced. A particular advantage, concluded by the author, was that the rubrics helped bridge multimodal pedagogy and assessment in the L2 classroom. Hafner and Ho (2020) aimed to assess scientific video documentaries produced by EAP students at a Hong Kong university. Stimulated interviews with seven teachers on what they looked for

and evaluated when rating the documentaries, demonstrated the need to adapt an existing rubric for ESP oral presentations to ensure multimodal aspects of the documentaries were represented, especially multimodal orchestration (combined effect of resources used), and to consider multimodal affordances.

Palmour (2024) conducted a large-scale study on multimodal oral presentation assessments in UK EAP programmes, to identify the theoretical, stated, perceived, and operationalized constructs underlying these. Tasks ranged from groupwork project presentations to individual presentations on students' topic of choice. Through questionnaires, interviews, and rating discussions, Palmour identified a mismatch between the predominantly language-focused criteria used in rating scales ("monomodal" constructs) and the multimodal resources that both students and teachers reported drawing on and valuing in oral presentation performances.

Rare examples of multimodal integrated formal assessment tasks outside the domain of university contexts, in large-scale standardised tests, are the computer-based EFL Listen-Write and Academic Listen-Speak tasks of the TOEFL Junior® test for 11–15-year-olds (So et al., 2015). Test-takers need to listen to an explanation accompanied by sequenced drawings, visualising key aspects and presenting phrases/words/numbers, on an (academic) topic and then synthesize or retell the information in writing (Listen-Write) or in speech (Listen-Speak). Kormos et al. (2020) showed that young learners displayed positive task-motivation towards these tasks, albeit finding Listen-Speak tasks more challenging and less enjoyable than Listen-Write ones. Michel et al.'s (2019) investigation of the Listen-Write tasks showed that learners performed well on these (similar to or better than on the test's independent-writing tasks), although there was a meaningful influence of Grade 7 learners' working memory on their Listen-Write performance, suggesting that efficient coordination of attentional processes might help young learners complete such multimodal integrated tasks. Overall, however, if we consider the naming of the tasks and scoring rubrics (see So et al., 2015), they might not fully represent a multi-modal integrated construct. The holistic rubrics draw raters' attention to the performance as a whole and contain mostly language-oriented descriptors similar to independent tasks with limited and generic reference to the multimodal input, which risks under-operationalising the multimodal integrated construct.

To reflect current-day language uses and broaden constructs represented in L2 formal assessments, we aimed to contribute to the emerging interest in multimodal L2 assessment. Specifically, we wanted to help extend the repertoire of tasks and rubrics beyond domains of language for specific purposes and contexts of alternative assessment (given that, as mentioned above, most L2 multimodal assessment work is currently situated in EAP/ESP and more formative assessment) and explore the affordance of multimodal tasks in general target language use domains (general L2 proficiency) and summative assessments. Additionally, we aimed to contribute to representing and operationalizing multimodal constructs in tasks and rubrics to a fuller extent than most existing formal assessments do. Thus, this article reports on a study in which we developed a type of multimodal integrated L2 test tasks and accompanying rating scales. While it is impossible to reflect the range of possible forms of multimodality in one study, as a first step in this line of research and test-task development, we decided to experiment with a task type that might be placed around the middle point of the multimodality continuum. In this early exploration of this topic area, we also thought it would be sensible to build on, rather than drastically deviate from, commonly used task types. Importantly, as our main focus

was on the assessment of language skills, we wanted to ensure that language modes (as conventionally defined) are robustly represented in the multimodal mix. We therefore opted for an integrated task including two language modes – aural and written – with further multimodality through visual semiotic resources commonly present in human communication – including gestures, body language, pictures, graphs, printed words/numbers, and spatial animations. As we explain below, we created a task type which we have called “viewing-to-write task.” These tasks require L2 learners to watch a video (visual-aural input) and produce a text (written output) based on the video. We decided to restrict the productive output to written mode only (rather than for example visual-written), anticipating the practicalities of many summative, timed proficiency testing contexts (as opposed to formative assessment contexts). As described below, we then investigated learners’ performance products and processes to gain insights into the tasks’ success in assessing this form of multimodal integrated language use.

RESEARCH QUESTIONS

To shed light on how English-L2 learners perform on multimodal integrated viewing-to-write test tasks, we designed two types of computer-based viewing-to-write tasks – *viewing-to-describe* and *viewing-to-compare-and-contrast* tasks. The language functions of describing, and comparing and contrasting, reflect authentic, prevalent functions in educational and professional settings. They are associated with distinct linguistic features (e.g., conjunctions, syntactic structures) and used in language learning and assessment tasks at various proficiency levels (though describing is often introduced at lower proficiency levels and comparing and contrasting at higher levels). The CEFR B1-C2 writing scale descriptors (Council of Europe, 2020) also illustrate ability differences in describing series of events and summarizing and synthesizing different points of view. Our first research question was therefore:

RQ1. How do English-L2 learners perform on *viewing-to-describe* test tasks and *viewing-to-compare-and-contrast* test tasks?

Since multimodal integrated viewing-to-write tasks currently do not commonly feature in formal language assessments, we additionally wanted to compare learners’ performance on these tasks with their scores on corresponding independent language-skills tests (listening and writing). Furthermore, we wanted to elucidate learners’ processing while completing the viewing-to-write tasks, to gain empirical insights into how successfully the tasks elicit the use of multimodal integrated skills. Additional research questions therefore were:

RQ2. How do learners’ scores on *viewing-to-describe* and *viewing-to-compare-and-contrast* test tasks relate to their independent listening and writing abilities?

RQ3. What forms of processing do learners employ when performing viewing-to-write test tasks?

METHODOLOGY

Instruments

Personal background questionnaire

Participant demographics such as gender, age, year of schooling, and home language(s) were gauged with a short background questionnaire, administered in Hungarian (participants' language of schooling and L1 for most) via Qualtrics.

Viewing-to-write tasks

Multimodal integrated viewing-to-write test tasks require learners to watch a recording including visual and oral input and then produce a short piece of writing based on the input. We developed two types of tasks:

- (a) *viewing-to-describe* tasks: learners watch a recording that explains orally and depicts visually and verbally (written words/phrases/numbers) how something is made, after which they write a short text describing the production process;
- (b) *viewing-to-compare-and-contrast* tasks: learners watch a videocast (a video recording similar to podcasts but also including visuals) with two experts discussing a specific topic, accompanied by visuals, after which the learners write a short report comparing and contrasting the experts' opinions.

We started the development process by drafting specifications for each task type, stipulating the intended construct and overall characteristics of the tasks as a whole, of the aural and visual aspects of the input, and of the expected output. Informed by pilot data (see below), minor revisions were made to some details of the specs. The final task specifications are available at <https://osf.io/xude7/>.

An important consideration was to design the tasks so that, in principle, it would be feasible for classroom teachers or testers in low-resource contexts to devise more versions. We therefore only used commonly available soft- and hardware (Word, PowerPoint, standard webcam, standard PC/laptop) and copyright-free materials (visuals from royalty-free image stock websites). We gained topic and content inspiration from websites, YouTube videos and media programmes on how to make things (*viewing-to-describe* tasks), and websites containing debating/group discussion ideas and media point-counterpoint articles (*viewing-to-compare-and-contrast* tasks). Topics for both task types were shortlisted based on the criteria: a) generally accessible subject while not overly common knowledge, b) rich in information points, c) possible to represent visually, and d) likely interesting to older adolescents and adults. Then, adhering to the task specifications, we developed task instructions, drafted scripts, and identified visuals (pictures, graphs, animations, words). While using existing online videos might be ideal from an authenticity point-of-view, we found that it is extremely difficult to find suitable ones for testing purposes – e.g., meeting all intended construct features, length and other practicalities, copyrights – and for parallel task versions. As we additionally know from prior

research that it is possible to craft authentic-sounding listening inputs (see e.g., Rossi & Brunfaut, 2021), we created the video inputs ourselves.¹

Linguistically, we targeted our scripts at the CEFR B- and C- levels. Content-wise, as set out in the task specifications and to be able to develop parallel task versions, we controlled for the number of main information points in each script, i.e. eight key production steps in each *viewing-to-describe* task and four main discussion points in each *viewing-to-compare-and-contrast* task. We estimated that this would be a suitable number of main content points for the input length (see below) and for target construct elicitation and allow for embedding within a larger discourse of more detailed information on each point. Topic-wise, for the *viewing-to-describe* tasks, it was important to select an object for which people generally do not know how it is produced, to avoid being able to successfully complete the task based on background knowledge. For the *viewing-to-compare-and-contrast* tasks, background knowledge was less of a concern, as the task requires to compare and contrast the views of the two specific speakers in the video, rather than of pros and cons on the topic in general.

The visuals in each task input were chosen to align with and complement the information that was orally conveyed. While most visuals were pictures that illustrated the main point being talked about in the relevant video frame, we included a range of additional visuals throughout each input. For example, key words which were specialised terminology that might be unknown were written out and visually linked to elements of the picture; animations reflected real-life movements described in the aural input; shapes drew attention to specific elements of the pictures focused on in the aural input; icons reinforced or added emotive information; graphs provided visualisations of or additional statistics and background information on what the speaker conveyed, etc. Importantly, the visualisations on their own were not so specific or revealing that they would allow for a successful task performance without listening to the video. At the same time, they intended to illustrate, enrich and further clarify the aural input, with the integrated aural-visual input enabling full comprehension of the video.

Then, we created voice-overs in PowerPoint. To showcase low-resource feasibility, the speakers in the recordings simply comprised the authors and willing colleagues, representing a mix of genders (two female; two male) and English-L1/L2 speakers (Australian English-L1, British English-L1, Dutch-L1, Hungarian-L1) who were experienced tertiary education teachers. They did not receive any training for this purpose, nor did they have acting backgrounds. For the *viewing-to-describe* tasks, we decided to have one speaker only explaining the production process, not showing the speaker in order to prioritise visual focus on how the product is made. This resembles the main features of for example many media programmes and online content on similar themes. For the *viewing-to-compare-and-contrast* tasks, we decided to show the speakers engaged in the discussion and to display visuals alongside them. This is similar to what one might see in many broadcast debates or interviews. As our speakers were not trained, any gestures or body language was simply their naturally occurring behaviour during the recording, as were any hesitations or fillers in

¹Note that when we were developing our tasks, the first AI tools were becoming publicly released. These could only handle and generate written text at that point and were in their infancy. Only after the completion of our project, AI tools became available that support audio and video generation. We are currently in the process of exploring the affordances of such new tools for task development.

their speech while they were reading out loud the scripts from a laptop screen outside of the camera view.

We opted for double play of the viewing input, without the possibility of pausing. Holzknacht and Harding (2023) provided an empirical justification for double play in the context of independent listening tasks. Given the quantity of information and length of input of our tasks (see below) and the additional demand of simultaneously processing multiple input modes, we hypothesised that double play would help minimise memory effects in multimodal integrated tasks. For similar reasons, we decided to allow test-takers to take notes on paper during the viewing stage. Then, once the voice-overs were recorded, we transformed the slides to a video format in PowerPoint and uploaded them to a private YouTube channel. For ease of standardised computer-based administration for our research purposes, we embedded the materials into Qualtrics. After the video viewing screen, test-takers moved on to a writing screen (including timer and word count displays), and composed their text relying on their notes.

We developed three *viewing-to-describe* tasks explaining how a) instant coffee, b) a train carriage, and c) contact lenses are made, and three *viewing-to-compare-and-contrast* tasks with speakers discussing the pros and cons of a) space tourism, b) zoos, and c) car usage. We first pre-piloted two tasks of each task type with eight experienced language teachers and testers (mix of English-L1 and English-L2 speakers). Based on their performances, their notes and follow-up focus groups we held with them, we made some small changes to the task instructions, writing time, word limits, and computer screen layout. Their overall feedback was highly positive. Then, we more formally piloted all six tasks with 31 EFL learners from our target population of older adolescents/adults (Hungarian-L1, EFL upper-secondary school students; $M_{\text{age}} = 17.2\text{y}$). Each learner completed three tasks, including at least one viewing-to-write and one viewing-to-compare-and-contrast task. Analysis of learners' viewing notes, written texts, and informal comments, and researcher observations suggested that all tasks and administration procedures worked well.

For the main study, we selected four tasks – two of each task type – given participant and time availability, resources, and envisaged statistical analyses. We proceeded with the *viewing-to-describe* tasks on how to make contact lenses (Describe-1) and a train carriage (Describe-2), and the *viewing-to-compare-and-contrast* tasks debating pros and cons of zoos (Compare&Contrast-1) and space tourism (Compare&Contrast-2), based on optimal discriminatory power and least risk of content familiarity. The *viewing-to-describe* tasks each convey eight content points on how something is made, explained in a 4.5-minute voiced-over video (instructions + single-play length) comprising visuals (pictures, animations, terms/phrases). Based on the input, learners need to write a coherent, 150-to-200-words magazine article describing the production steps. The *viewing-to-compare-and-contrast* tasks each comprise a videocast introduced by a host, with two experts expressing a mix of shared, opposing and no stated views on a topic (covering four issues). The 4.5-minute video (instructions + single-play length) shows the speakers interacting and visuals (pictures, animations, graphs, terms). Learners need to write a 200-to-250-word report comparing and contrasting the speakers' views (summarize key points, clearly indicating similarities and differences between speakers' views). The tasks are available at <https://osf.io/xude7/>. An example of each task type with stills from the videos are shown in Figure 1.

Viewing-to-describe task

1. Task instructions

TASK
Describe a process

You will watch a video describing how something is made. The video will be played twice. You are allowed to take notes while watching.

Then, write a short magazine article describing the different steps in the production process. Your text should be coherent and 150-200 words in length. A title will be provided for you.

The video will start now.

2. Viewing (watching video)



3. Writing

1800

[time counter]

Now write a short magazine article describing the different steps in the production process. Your text should be coherent and 150-200 words in length. You have 18 minutes. A title is provided for you.

0/200 words [word counter]

How a train carriage is made

Viewing-to-compare-and-contrast task

1. Task instructions

TASK
Compare and contrast

You will watch a video in which two people discuss their views on a certain topic. The video will be played twice. You are allowed to take notes while watching.

Then, write a short report comparing and contrasting the two speakers' views. Summarize the key points discussed, indicating clearly any similarities and differences between the two speakers' views. Always make clear whose views you are reporting.

Your text should be coherent and 200-250 words in length. A title will be provided for you.

The video will start now.

2. Viewing (watching video)



3. Writing

2000

[time counter]

Now write a short report comparing and contrasting the two speakers' views. Summarize the key points discussed, indicating clearly any similarities and differences between the two speakers' views. Always make clear whose views you are reporting. Your text should be coherent and 200-250 words in length. You have 20 minutes. A title is provided for you.

0/250 words [word counter]

Commercial space travel: the way to go?

Figure 1. Viewing-to-write task examples (Train image by benfuenfundachtzig @pixabay.com).

Language proficiency measures

For RQ2, to measure learners' independent skills abilities, we administered the listening and writing tests (including grammar and vocabulary tests) of the Aptis General test (see <https://www.britishcouncil.org/exam/aptis> for more information and sample tasks), which reports CEFR levels. The General version was used as our viewing-to-write tasks are aimed at older adolescents and adults, and our participants were anticipated to be 16–19-year-olds.² The listening test consists of multiple-choice and matching items which assess listening for specific information (dates, names, etc., and factual information and identifying/inferring speaker attitude/intention/opinion; the writing test consists of four holistically-scored tasks assessing word/phrase writing, personal information sentence writing, interactive question-answer writing, and email-response writing (O'Sullivan et al., 2020). Piloting with 26 learners suggested that Aptis General was suitable for the main study.

Post-task interview protocol

To examine learners' processing of viewing-to-write test tasks (RQ3), we conducted post-task recall interviews whereby the viewing input and participants' notes acted as stimuli to recall one's viewing, comprehension and production processes after each task. Interviews were conducted individually, following a standardized protocol and in participants' choice of language (Hungarian in all cases). The protocol included questions the researcher could use to elicit reports of thought processes on viewing and note-taking (e.g. "What was on your mind while watching this?", "What helped you understand this part of the video?") and on the writing phase (e.g. "How did you start your writing?", "Did you look at your notes

²The Aptis General Guide states that this test "is suitable for most test-takers" (British Council, 2023, p. 2). Aptis Teens, in contrast, is designed for 13–17-year-olds (British Council, 2023) and was thought to be potentially less motivating for the upper age range of our participants. Pilot participant feedback indicated that the Aptis General content was suitable and accessible to our population.

while writing?”). After a participant completed a task, the viewing input was replayed and the researcher paused it at slide transition points. The learner was then prompted to recall and verbalise their processing. If not already shared, participants were additionally prompted to describe their notetaking and writing processes. Piloting with three learners, on two tasks (one *viewing-to-describe* and one *viewing-to-compare-and-contrast*), indicated that the methodology and procedures were effective. Interviews were audio-recorded, automatically transcribed, and manually corrected where needed, translated into English and double-checked by two Hungarian-L1 EFL specialists.

Participants

One hundred and thirty-four EFL learners from six upper-secondary schools in Hungary participated in our study.³ They were in Grades 10 (49%), 11 (20%) or 12 (31%). Their ages ranged between 15 (1.5%) and 19 (6%) ($M = 16.64y$, $SD = .969$). Sixty percent were male; 40% female. Ninety-eight percent had Hungarian as their home language (two spoke Chinese at home and one Arabic). Based on the Aptis results, their English proficiency ranged between CEFR A2-C levels (listening: A2 = 1%, B1 = 2%, B2 = 19%, C = 78%; writing: B1 = 5%, B2 = 46%, C = 49%).

Data collection procedures

Ethics approval was granted by Lancaster University’s FASS-LUMS Research Ethics Committee. Prior to data collection, learners and their parents were provided with information sheets in Hungarian and written consent was sought. Permission was also obtained from participating schools and teachers. Participation was voluntary.

Data were collected in two sessions during school hours, in classrooms, using PCs/laptops and ear/headphones. During the first session, participants completed the Aptis tests. During the second session, participants completed the bio-data questionnaire and viewing-to-write tasks. Twenty participants conducted the second session individually, additionally doing post-task interviews.

Given the available time for the second session (1.5 hours) and the required effort, the main study participants who were not involved in the interviews each completed three tasks,

Table 1. Counterbalanced design for viewing-to-write data collection.

Group		First task	Second task	Third task
Group 1	Group 1a	Describe-1	Compare&Contrast-1	Compare&Contrast-2
	Group 1b	Describe-1	Compare&Contrast-2	Compare&Contrast-1
Group 2	Group 2a	Describe-2	Compare&Contrast-1	Compare&Contrast-2
	Group 2b	Describe-2	Compare&Contrast-2	Compare&Contrast-1
Group 3		Describe-1	Describe-2	Compare&Contrast-1
Group 4		Describe-2	Describe-1	Compare&Contrast-2

³Twenty-four pilot participants’ data was incorporated into the main study, as they were part of the same population, completed all instruments used in the main study, and no changes were made to instruments and procedures between pilot and main study.

Table 2. Counterbalanced design for post-task recall interviews.

	First task	Recall	Second task	Recall
Order A	Describe-1		Compare&Contrast-1	
Order B	Describe-1		Compare&Contrast-2	
Order C	Describe-2		Compare&Contrast-1	
Order D	Describe-2		Compare&Contrast-2	

including at least one *viewing-to-describe* and one *viewing-to-compare-and-contrast* task.⁴ Table 1 specifies the counterbalanced design that was employed, controlling for task and order, with participants randomly assigned to groups. Participants always started with a *viewing-to-describe* task since the pilot study suggested they were more familiar with this language function.

The 20 participants volunteering for post-task recall interviews each completed two tasks, one *viewing-to-describe* and one *viewing-to-compare-and-contrast* task, given the extra time needed for process verbalisations. Table 2 shows the counterbalanced design adopted, controlling for task and order, with participants randomly assigned to groups.

Rating scales

To evaluate learners' performances, we developed two analytic rating scales – one for rating *viewing-to-describe* and one for *viewing-to-compare-and-contrast* performances. Based on a combination of our expertise in scale development for proficiency testing, familiarity with rating scales employed by various exam organisations, and theoretical insights into the *viewing-to-write* construct, we determined four rating criteria and developed descriptors for five performance bands. We kept in mind our aim to develop usable, uncomplicated instruments for classroom teachers, which are simultaneously rich enough to provide meaningful insights into learners' abilities, and reflect criteria commonly reported in education systems.

Specifically, to assess learners' comprehension of the video input, we established a *Viewing for writing* rating criterion. Given the difference in nature of the two task types and underlying constructs (describe; compare-and-contrast), this criterion was operationalised differently for the two task types. We anticipated that such task-type specificity would offer more insightful information for classroom teachers and learners than a vague, language function-independent criterion. Additionally, given each task version's distinct topic, we developed crib sheets to assist the raters and ensure rating consistency. These specify the necessary content points from each *viewing-to-write* input that need to be represented in a written performance.

To evaluate learners' written production for important aspects of writing ability within the context of the video input, we established three rating criteria in addition to *Viewing for writing*, namely *Organisation and Structure*, *Language Use*, and *Mechanics*. The *Language Use* and *Mechanics* criteria were kept the same for the two task types. Given the different language functions (describing; comparing-and-contrasting), the *Organisation and Structure* criterion was operationalised partly differently for the two task types. As

⁴While the 24 pilot participants integrated into the main study completed three tasks, data from only two was used in the main study, as the third task was a version not selected for the main study.

a starting point, part of our descriptors for these three criteria were inspired by a scale developed for another project by the second author and by the reading-into-writing scale from Trinity College London's Integrated Skills of English exam (<https://www.trinitycollege.com/qualifications/english-language/ISE/ISE-results-and-certificates/ISE-rating-scales>); the latter also informed our scales' layout and formatting). Additionally, the descriptors were shaped by our expertise with rating scales and the characteristics of the target construct. Importantly, we included references in the descriptors to the video to ensure that the multimodal integrated construct was appropriately represented in these writing-oriented criteria. For example, the *Language Use* criterion included the level-4 descriptor "Vocab: [...] wide range and high sophistication commensurate to the video input". Based on this descriptor, the rater had to consider lexical aspects of the written performance in light of the lexis offered in the video, and, to award a top score of 4, the learner's text had to include the range of lexis from the video, and possibly even go beyond that of the video (both its aural and written input). Another example is that the *Organisation and Structure* criterion of the Viewing-to-compare-and-contrast scale contained the level-1 descriptor "Limited or poor signposting; [...] several ambiguous or missing referents to signal an individual speaker". This explicitly required evidence about the test-taker's (lack of) ability to establish which of the two speakers presented which content point in the video input and accordingly use the correct referent to signpost it.

The *Viewing for writing*, *Organisation and Structure*, and *Language Use* criteria were rated on a scale of 0–4, whereas *Mechanics* was rated on a scale of 0–2 to avoid comparatively overrepresenting this more technical aspect of writing.

As a first step, we pre-piloted the rating scales with nine experienced language testers (one English-L1, eight English-L2 raters; nine for viewing-to-describe, three for viewing-to-compare-and-contrast) on 10 pilot study performances (five viewing-to-describe and five viewing-to-compare-and-contrast performances covering a diverse quality range based on our own screening). Analysis of the pre-pilot raters' scores, their rating notes and data from a follow-up focus group interview suggested that the scales and crib sheets required only minor changes. The minimally revised scales were then formally piloted by another English-L1 rater on all pilot study performances (81). The (pre)piloting indicated that the scales were detailed and user-friendly, represented a suitable and well-balanced range of (sub) constructs, and scores reflected the quality of multimodal integrated performance. Appendix A shows the rating scales. The crib sheets (and scales) are available at <https://osf.io/xude7/>.

The pilot rating data also supported the viewing-to-write tasks' quality; descriptive statistics on the performance scores showed that the tasks and rating scales were able to distinguish performance levels across the criteria, and that the task versions in each task-type set (three *viewing-to-describe*; three *viewing-to-compare-and-contrast*) were similar in difficulty.

Analyses

First, participants' viewing-to-write task performances were scored using the two rating scales. This was done by two experienced raters (one female English-L1 and one male English-L2 speaker) who were familiarised with the tasks, rating scales, and crib sheets, and trained on 30 performances, achieving good inter-rater reliability (*viewing-to-describe*

Spearman's $\rho = .94$, $p < .01$; *viewing-to-compare-and-contrast* Spearman's $\rho = .77$, $p < .01$; [Appendix B](#) shows the coefficients for the individual rating criteria). The two raters' scores were then averaged for further analyses. Second, participants' Aptis scores were provided by the British Council, with listening, vocabulary and grammar scored automatically and writing by regular Aptis raters (Dunn, 2020).

To answer RQ1, we calculated descriptive statistics for the scores on the individual rating criteria (*Viewing for Writing*; *Organisation and Structure*; *Language Use*; *Mechanics*). We also created two composite scores: a) the total score on each task, and b) one that combined the scores on the three writing-related criteria only (*Organisation and Structure*, *Language Use*, *Mechanics*). Additionally, we conducted Pearson correlation analyses, for each task version, examining the inter-relationship of scores on the four rating criteria and with the composite scores.

For RQ2, we ran descriptive statistics on the Aptis listening and writing scores and carried out correlation analyses to establish any links between the Aptis and viewing-to-write task scores (*Viewing for Writing* criterion; composite writing score; Total score). Pearson correlations between the independent skill scores – Aptis writing and listening – were also calculated. For effect size benchmarks we used Plonsky and Oswald's (2014) criteria: small (0.40), medium (0.70), large (1.00).

Then, to fully examine the interrelationship between learners' scores on viewing-to-write tasks and their independent listening and writing abilities, and the potential moderating role of task-type, we used mixed-effects modelling. This analysis comprised 357 viewing-to-write observations (134 students each completing two/three viewing-to-write tasks). We considered the viewing-to-write task total scores as interval variables. As the dependent variables were normally distributed, we utilized generalized linear mixed-effects models (GLMMs), using *lmer* function of the *lme4* package (version 1.1.27.1; Bates et al., 2015) in R (version 4.1.2; R Core Team, 2021). The significance of fixed effects was assessed with the Satterthwaite approximation for degrees of freedom using the *lmerTest* package (Kuznetsova et al., 2017). The model included random intercepts of Participant and Task content. We initially considered the random slopes of Participant and Task; however, the maximal model (Barr et al., 2013) failed to converge; consequently, the random slopes were removed. The final code for our model on predictors of viewing-to-write task total scores was:

$$TotalViewing\text{-}to\text{-}WriteTask \sim AptisListening + AptisWriting + Tasktype \ (1|Participant) + (1|Taskcontent)$$

To establish learners' processing while completing the viewing-to-write tasks (RQ3), we analysed the English transcripts of the post-task recall interviews, using the qualitative data analysis software Atlas.ti. We developed a coding scheme comprising seven main codes, informed by the nature of the viewing-to-write tasks and general skills involved. Three codes focused on independent skill aspects: a) listening, b) visual viewing, and c) writing; four codes focused on integrated aspects: a) integration of audio and visual input, b) integration of input into notes, c) integration of input into writing, and d) integration of notes into writing. The coding scheme was first tried out on four transcripts (one from each task) and re-applied again a month later. This indicated that it allowed for comprehensive and replicable capturing of processing (96% intra-coder agreement). Consequently, the coding scheme was employed to analyse all interviews.

RESULTS

With respect to RQ1, the descriptive statistics of the four viewing-to-write tasks (Appendix C) indicated that overall participants performed similarly on the four tasks. The total scores on the four tasks revealed that students achieved on average 65%-69% of the maximum possible score of 14. The tasks also had similar minimum and maximum scores and standard deviations.

The tasks' *Viewing for Writing* criterion scores showed that participants rendered on average 63%-67% of the information units conveyed through the multimodal input. The mean scores for the writing-composite total (summing *Organisation and Structure*, *Language Use*, *Mechanics*) fell within a similar percentage range, 64%-69%. Although all criteria's scores on each task were negatively skewed, kurtosis values for the total scores did not suggest too narrow clustering, with particularly good spread observed in the *viewing-to-describe* task on contact lenses and the *viewing-to-compare-and-contrast* task on space tourism. The writing-composite scores equally indicated an appropriate spread of scores.

While all tasks performed quite similarly statistically, the *viewing-to-compare-and-contrast* task on zoos was just slightly easier ($M = 68.6\%$). The zoo topic is likely to have been the most accessible and familiar of all four, including in terms of vocabulary associated with it, as demonstrated by the somewhat higher mean *Language Use* (71.5%) and *Mechanics* (66.0%) scores on this task. Even so, the pilot study had confirmed that it was necessary to watch and understand the input to successfully complete this task. The *viewing-to-describe* task on trains ($M = 64.6\%$) was minimally more challenging compared to the other tasks, with the difficulty seemingly situated in the *Language Use* criterion ($M = 62.0\%$); indeed, this task's input included a few more technical terms (e.g., welded, bolsters), although some were written out in the multimodal input (e.g., underframe).

Table 3 shows correlations between the individual rating criteria, and also with the total scores and writing-composite scores. Except for the link between the *Mechanics* scores versus the *Viewing for Writing* and *Organisation and Structure* scores, all correlations were above .5, indicating medium to large effect sizes. The fact that *Mechanics*, representing spelling accuracy and punctuation control, correlated only moderately with those two criteria is not unexpected given their focus on content ideas and textual flow. The stronger correlation between *Mechanics* and *Language Use*, however, is expected, given the inherent connection between lexical knowledge and spelling, and between syntactic knowledge and punctuation in writing.

The overall strength of correlations warranted conducting factor analysis, using the four rating criteria's scores. The Kaiser-Meyer-Olkin value (.755) exceeded the .6 recommended value and Bartlett's Test of Sphericity reached statistical significance, thus supporting factorability of the correlation matrix. Principal component analysis revealed that the four criterion scores loaded on a single factor with an eigenvalue of 2.761 explaining 69.01% of the variance, suggesting that the four criteria tap into a single underlying construct.

Regarding RQ2, the descriptive statistics (Appendix D) showed that participants achieved relatively high scores on the Aptis Listening test ($M = 42.30$; max. possible score of 50). Although there was a fair spread of listening scores ($SD = 4.86$), the kurtosis value (4.776) indicated that these scores were generally quite closely clustered. Modulated by

Table 3. Correlations of the rating criteria and total scores of the viewing-to-write tasks.

	Viewing for writing	Organization & structure	Language use	Mechanics	Writing composite
<i>All four tasks (n_{participants} = 134; n_{performances} = 357)</i>					
Total score	.889**	.908**	.875**	.596**	.961**
Viewing for writing		.754**	.646**	.340**	.728**
Organization & structure			.728**	.431**	.905**
Language use				.568**	.920**
Mechanics					.688**
<i>Describe task: Contact Lenses (n = 93)</i>					
Total score	.875**	.888**	.894**	.572**	.950**
Viewing for writing		.693**	.643**	.298**	.681**
Organization & structure			.764**	.389**	.898**
Language use				.566**	.940**
Mechanics					.675**
<i>Describe task: Train (n = 88)</i>					
Total score	.869**	.905**	.889**	.616**	.966**
Viewing for writing		.726**	.642**	.346**	.711**
Organization & structure			.734**	.443**	.906**
Language use				.590**	.926**
Mechanics					.694**
<i>Compare & Contrast task: Zoo (n = 91)</i>					
Total score	.912**	.911**	.867**	.567**	.961**
Viewing for writing		.784**	.683**	.331**	.764**
Organization & structure			.712**	.426**	.908**
Language use				.524**	.905**
Mechanics					.671**
<i>Compare & Contrast task: Space Tourism (n = 85)</i>					
Total score	.913**	.941**	.882**	.617**	.972**
Viewing for writing		.826**	.698**	.398**	.791**
Organization & structure			.778**	.487**	.933**
Language use				.562**	.920**
Mechanics					.696**

**Correlation is significant at the 0.01 level (2-tailed).

participants' Aptis vocabulary and grammar scores, this meant that the average listening proficiency fell in the CEFR C levels, based on Aptis-CEFR calibration reporting (O'Sullivan, 2015). Participants also performed well on the Aptis Writing test ($M = 44.13$; max. possible score of 50), with scores relatively narrowly clustered ($SD = 3.56$). According to the Aptis-CEFR calibration reporting and modulated by participants' Aptis vocabulary and grammar scores, the average writing proficiency fell at the upper end of B2.

The correlation between the Aptis listening and writing tests was $r = .539$ ($p < .001$), indicating that participants' listening and writing skills were associated. Nonetheless, the strength of association did not reach the threshold at which inter collinearity presents a problem in subsequent linear mixed effects modelling.

Table 4 examines the relationship of the *Viewing for Writing* criterion scores, the writing-composite scores, and the total task performances with participants' performance on the Aptis listening and writing tests. The correlation values show consistently moderate (mainly for the *Viewing for Writing* criterion) and large (for the writing-composite and total scores) effect sizes. Additionally, viewing-to-write task scores (*Viewing for Writing* criterion, writing-composite, and total scores) are consistently more strongly linked to the Aptis Listening than Aptis Writing scores.

The results of the general linear mixed effects modelling in Table 5 indicate that both the Aptis listening and writing scores were significant predictors of the total score participants

Table 4. Correlation between viewing-to-write task performance and Aptis listening and writing scores.

	Aptis Listening	Aptis Writing
<i>All four tasks (n_{participants} = 134; n_{performances} = 357)</i>		
Viewing for writing	.487**	.393**
Writing-composite score	.606**	.533**
Total score	.601**	.515**
<i>Describe task: Contact Lenses (n = 93)</i>		
Viewing for writing	.425**	.393**
Writing-composite score	.659**	.549**
Total score	.616**	.530**
<i>Describe task: Train (n = 88)</i>		
Viewing for writing	.528**	.466**
Writing-composite score	.612**	.592**
Total score	.626**	.588**
<i>Compare & Contrast task: Zoo (n = 91)</i>		
Viewing for writing	.536**	.327**
Writing-composite score	.624**	.441**
Total score	.625**	.420**
<i>Compare & Contrast task: Space Tourism (n = 85)</i>		
Viewing for writing	.485**	.400**
Writing-composite score	.659**	.549**
Total score	.616**	.530**

**Correlation is significant at the 0.01 level (2-tailed).

Table 5. Summary of the linear mixed effects model of the predictors of the total viewing-to-write task scores.

Fixed effects	Estimate	Standard error	Standardized β	t	p
Intercept	−7.79	1.73	9.26	−4.49	***
Aptis Listening	.22	.03	2.14	6.91	***
Aptis Writing	.18	.04	1.27	3.91	***
Task type	−.26	.20	−0.27	−1.26	
Random intercept	Variance	SD			
Participant	1.82	1.35			
Task content	.01	.16			
Log likelihood	−673.78				
AIC	1361.6				
Marginal R ²	0.411				
Conditional R ²	0.737				

Note. AIC = Akaike Information Criterion.

*** $p < .001$.

achieved on the viewing-to-write tasks. Task type (describe; compare-and-contrast) did not exert a significant effect on performance scores. The standardized beta values show that with one SD increase in the Aptis Listening score the total viewing-to-write score increases by 2.14 SD, and with one SD increase in the Aptis Writing score the total viewing-to-write score increases by 1.27 SD. The variance explained by the fixed-effects predictor (Marginal $R^2 = .411$) suggests a medium-size effect (Plonsky & Ghanbar, 2018).

Concerning RQ3, the post-task recall interviews showed that the 20 participants' processing of the viewing-to-write tasks involved aspects of listening, watching, writing, as well as bridging these skills by integrating aural and visual information, transforming viewing input into notes or directly into their written performance, and integrating their notes into their written performance. Appendix E reports the number of times a form of

processing was evidenced in the interviews. Fairly similar proportions across the two task types and the four task versions were found. The most frequently reported processes concerned aspects of listening (30%), writing (21%), and integrating notes into one's writing (21%). Overall, 41% of the codes indicated integration of different skills and information. The interviews' main purpose, however, was to shed light on the nature of the task processing, which is best gauged through participants' own words.

Visual processing and integration

A first question to establish the multimodal nature of the input is whether learners actually looked at the video. All participants referenced visual aspects of the task input in their recalls. Some simply acknowledged it in general terms: *"I saw the picture, of course"* (Describe-ContactLenses-S09), *"I watched the two people talking"* (CompareContrast-Zoo-S01), or *"They were rotating the picture and it was so attention-grabbing"* (Describe-Train-S20). A few explicitly stated that the visuals helped them to remember content that would be useful for their writing. For example, for the *viewing-to-compare-and-contrast* task on zoos, one participant formed an associative memory between individual speakers (who could be heard and seen in the video) and pictures to match the right opinion to the right speaker for their writing:

What made it easier was the pictures and the video. They helped me remember which one of the speakers was talking. For example, with the panda picture, I remembered who talked about it. (CompareContrast-Zoo-S19)

Only one participant seemed to have consciously deprioritised the visual input at some point, while nevertheless referencing visuals elsewhere in their interview.

I didn't pay that much attention to the video, I paid more attention to the audio. I tried to understand it based on the sound, which I did, and I tried to write down as much information as I could. (Describe-ContactLenses-S13)

Some found the simultaneous processing of multiple channels (audio, pictures, speakers, content, note-taking) somewhat challenging, confirming our hypothesis on the cognitive demands of these tasks. Thus, since the input would be played twice, they strategically balanced their attention differently between the various channels on each play. For example, some prioritised watching and listening during the first play and note-taking during the second, while others did the reverse⁵

On the first watching, I really paid attention so that I could see everything, hear the text and see the pictures. Then I was sure I understood everything and the second time I was more focused on taking notes. (Describe-Train-S03)

On the second listening, I paid more attention to the video, because the vast majority of my notes were already written down. [during the first listening] (CompareContrast-Zoo-S09)

Significantly, three-quarters of participants explicitly indicated how combining the visuals with the aural channel aided their understanding. For example, S15 explained how

⁵Many other participants did a bit of everything during both plays.:

a written-out term for an object, a picture and an animation made that step of train manufacturing described in the audio genuinely clear:

I didn't know this word "bolster," so it was good to see it [in the picture]. Also, when they said, for example, that it should be turned, the picture turned. And I could really see what it was, the picture was there. Without the pictures, I couldn't have imagined what it was and where it was.

Or, regarding the *viewing-to-compare-and-contrast* tasks which include a video of the speakers interacting, when the researcher asked what helped understand the input, S04 mentioned integrating aural content with speaker gesticulation:

That the video had pictures and well the way they spoke so clearly. And the hand gestures, their body language, too. For example, the lady, when she was saying how wonderful the view must be, she was pointing with her hands. (CompareContrast-SpaceTourism-S04)

In another example, S20 explained how combining the aural input with a graph gave them certainty on a speaker's position in the space tourism task:

I didn't know if the speaker was regarding tourism as a positive or negative, because it's a little bit of both. [...] And then the speaker ends up saying it as positive, that this is "boosting" the "industry". Here I looked up to the picture to see if that would help, I looked at the graph. It's a pretty easy graph, you can totally see what it's talking about, how much the "adventure" has been boosted. So it was clear afterwards that the speaker was mentioning it as a positive thing. (CompareContrast-SpaceTourism-S20)

Listening processing and integration

Participants also evidenced a wide variety of listening processes, ranging from low-level decoding to high-level text representation in function of the writing tasks. For instance, participants' reports made aspects of lower-level phonological and lexical decoding visible and attention to pace and enunciation, e.g.

Phonological decoding: I heard the word "welded" but I didn't know the word. And then I wrote down a "w" and I tried letters. (Describe-Train-S20)

Lexical decoding: I wrote down that you need "epoxy paint" and "paint gun" to do it. So it's not a machine that does it, it's people. (Describe-Train-S03)

Pace and enunciation: They both speak very slowly, very articulated, very beautifully, so that helps a lot (CompareContrast-SpaceTourism-S16);

Here I noticed they spoke a bit faster, but I didn't have any problems with it. (CompareContrast-Zoo-S11)

Their comments also evidenced inferencing processes regarding lexis and content, e.g.

Lexis: I didn't know the word "sterilise" [...] I couldn't guess from the word itself, but then he explained it means that they check for bacteria or something. (Describe-ContactLenses-S06)

Content: I think in the last part, about the door, they didn't say how it was built into the carriage. They may have thought it was obvious that it was built into the panel there, but it was a bit short on information. (Describe-Train-S12)

Examples of higher-level processing include characterisations of the input's organisation and genre. Organisation-focused comments also demonstrated awareness of the writing task that would follow (manufacturing description; comparing/contrasting views), with participants directing their attention accordingly, e.g.

Organisation: The video told me in a logical order what a carriage was made of. It was actually pretty good that it went in order, and there wasn't too much information about any one thing, so it was relatively clear to me what I thought was important to know and what wasn't. (Describe-Train-S12)

Genre: The conversational nature of it helped. (CompareContrast-SpaceTourism-S18)

Furthermore, participants' comprehensive meaning representation and engagement with the input transpired from the connections they made between the content and their own background knowledge, elaborating on it and sharing their own reactions, e.g.

I was a little shocked that there are over 140 million people in the world who wear [contact lenses], which is a pretty big number. But when you think about it, it's understandable, because more people than that wear glasses. (Describe-ContactLenses-S09)

Notetaking sometimes aided reaching full text representation, e.g.

After I listened to it once, in a way that I fully paid attention to it and also made notes, it was easier to put [the overall meaning] together. (CompareContrast-Zoo-S01)

Participants additionally described forms of metacognitive processing, e.g. comprehension monitoring:

The man was talking mainly, and so I tried to understand it at the first listening, and then I thought something like, well, this must be what it's about. And then on the second listening, when I was concentrating on what the man was saying, I understood that no, actually, that is what it was about. (CompareContrast-Zoo-S17)

Another type of metacognitive processing was task-oriented viewing (watching/listening/notetaking), integrating viewing and writing processes, as participants knew what kind of writing they would need to do next, e.g.

I was just writing [notes while viewing], I was thinking about the fun fact, I wanted to put that in [my notes] because it's an article [I need to write] and I did not want to be very dry. (Describe-Train-S08)

It was also very apparent that notetaking while processing the video both served as a content memory aid and enabled participants to make transitions from input comprehension to writing. For example, it helped to select and identify relevant and sufficient information from the input:

I weighed when taking notes about which parts were about putting the train together and to include most of that. Because that was the task. (Describe-Train-S16)

Notetaking also helped establish coherence between points for writing purposes:

For the first [listening], I tried to write down specific words, [...] phrases that I would like to copy into my essay. And on the second [listening], I tended to link them, I wrote conjunctions, because it's a great help if you have complete sentences that you can just copy into the essay. [...] I drew arrows to connect things. (Describe-Train-S20)

Additionally, participants used notetaking as a tool for structuring and further processing of what they understood from the multimodal input in function of what they needed to write. Distinct approaches were employed for the two target language functions, with many participants focussing on sequencing content points for the descriptive writing tasks, e.g.

I numbered the production process. Step one, step two and so on. (Describe-ContactLenses-S13);

I keep drawing arrows, marking how things happened in order. (Describe-ContactLenses-S02)

For the comparative writing tasks, many participants visually separated the two speakers' views in their notes as they were processing the input, labelling the nature of each expressed opinion and indicating how it related to the other speaker's opinion, because "*it was much easier to write the essay based on that*" (Describe-Zoo-S13), e.g.

I wrote down the two [speakers'] names: one on the left and one on the right, and I wrote their opinions next to each of them. Somewhere I wrote that, for example, they share the same opinion, but somewhere I wrote that it wasn't quite the same. (Describe-Zoo-S15)

Writing processing and integration

Participants indicated attention to lower- and higher- order writing processes and regulation mechanisms. At a lower level, for example, they considered the linguistic accuracy of their writing:

"by" or "with": I was thinking about which one is correct. "By hand" and then "with a paint gun". (Describe-Train-S04);

I spelled "followed by" as one with one L and two O's. And it was obviously a typing error [...] and I could sort of correct it. (Describe-Train-S16)

Ensuring lexical sophistication and diversity was also a consideration, with participants purposefully enriching their linguistic choices and avoiding word repetitions, e.g.

In a sentence I wrote "good" twice. [...] I said [...] it's a good programme for small children. So it can be exciting. And then I put "exciting" instead of "good", to make it more diverse. (CompareContrast-Zoo-S11)

Furthermore, giving that the participants were L2 learners, it was unsurprising that they sometimes adopted compensation strategies for lexical knowledge gaps, e.g.

I thought of a word, but I didn't know how it was in English, I had to rewrite the sentence a little bit to get my thoughts through, but not with that word specifically, but a little bit differently. (Describe-Train-S07)

At a higher level, participants explained their planning processes before and while writing, e.g.

I tried to get my thoughts together roughly and what ideas I had, and how I wanted to write. So I like to put it together in my head. (Describe-ContactLenses-S10)

There was an instance where I stopped and thought about how to continue, how should I phrase it. (CompareContrast-SpaceTourism-S04)

Participants also conveyed how they had organised their writing. Many mentioned sandwiching the main content between a brief introduction and conclusion. For the describe tasks, their intention was for the middle part to systematically reflect the manufacturing steps from the input, e.g.

First I thought I would do an introduction. I wrote that there were only glasses until the 70's [...]. And then I used a question to lead into the part where it says how they are made [...]: "how are they made?". And then I switched from that to the part – after I finished describing how they are made – the part about quality control, sterilisation and then packaging and delivery. And as a finishing sentence, I wrote that it's a great thing to have an alternative to glasses, to be able to do sports activities without glasses. (Describe-ContactLenses-S09)

For the compare-and-contrast tasks, different structuring approaches were noted for the main body, e.g. reflecting speaker alternations, separating positive from negative views, or organising according to thematic points. e.g.

I wrote the positive parts of the whole thing first, and then I started writing the negative, the disadvantages of the whole thing. (CompareContrast-SpaceTourism-S14)

I tried to divide it into three main bullet points or clash points. One I think was that they agreed with the beauty of this and that this could make tourism move forward. Or the two clashes are how much this will advance science and how much of a scientific purpose it serves. And whether or not it will raise awareness about the environment. But I've given titles to everything that will be in it. I worked with short paragraphs. (CompareContrast-SpaceTourism-S06)

When discussing written organisation, integration of different channels – including the visuals – transpired from participants' statements, for example how the nature of the input had helped them structure their writing:

All I would note here, and this will appear in the essay, that you can see quite clearly where the ends of the slides are. So in light of that, I was able to structure the paragraphs better, here a unit of thought or a step in its assembly was pretty clearly concluded. (Describe-Train-S16)

Or, how they drew on their notes during writing:

I wrote it all from my notes. I incorporated the key words and added my own words. (Describe-ContactLenses-S18)

Or a combination of the multimodal input (e.g., video frames) and their notes:

If you looked at the screen you could see what you could expect from that particular frame. [...] I could structure it and that was a big help. It helped me in my essay too. Because of my notes I could write different paragraphs. (Describe-Train-S20)

To achieve a well-organised text for the targeted language function, participants furthermore made use of cohesive devices, e.g.

I opened with "first and foremost" and "secondly" and "lastly" to make the steps and the process understandable to the reader. (Describe-ContactLenses-S14)

Linking words, so I tried to use like "he thought this", "on the contrary", "they both agreed on". (CompareContrast-SpaceTourism-S16)

Participants also aimed to reflect in their writing the meaning representation they had built from the multimodal input. However, where their comprehension was incomplete, this affected their writing, including its coherence, e.g.

I can express myself in English, it's just the parts [of the input] that were confusing, I didn't know how to write them in a way that [...] the reader can understand. [...] What I noted down and understood, I could put into words quite easily. I wrote in "in addition" and things like that to make the text flow, but the parts that I didn't understand, I felt a little bit that I didn't know what I was writing. (Describe-Train-S20)

A few participants displayed how genre awareness (article; report) had influenced their writing, e.g.

I had time to think it through, to make sure to use the parameters and stylistic features required for the article. So from that point of view, I tried to focus on the tone of the English articles I read. (Describe-Train-S16)

Finally, participants' comments suggested metacognitive monitoring of their writing, and editing where necessary, e.g.

When I had already written down a part of the text, I looked at it a little bit, changed my mind, erased it, and wrote it a little bit differently because I didn't quite like the way it was there the first time. (Describe-ContactLenses-S09)

DISCUSSION AND PRACTICAL IMPLICATIONS

Our research shows that the type of viewing-to-write tasks developed in our study allow higher proficiency EFL learners to demonstrate their L2 skills at their level; with 64%-69% average scores and reasonable spread, they are neither too easy nor difficult and able to discriminate among learners with different abilities. The rather similar statistics across our four task versions furthermore indicate that it is possible to construct multiple well-functioning tasks. At the same time, some variability in difficulty can be realised, as shown by the slightly easier zoo task and slightly more difficult train carriages task. In both cases, general linguistic familiarity (or lack thereof) related to the input (broader topic domain) seemed to be the modulating factor, given the differential achievement on the *Language Use* criterion versus other criteria and task versions. The qualitative interview data – demonstrating focused task processing and suitable engagement without undue anxiety or concerns over difficulty/ease – also support that the tasks were targeted at the right level. Additionally, our findings suggest that these multimodal integrated tasks are appropriate for older-adolescent and adult populations. Whereas Michel et al. (2019) found a slight risk of cognitive individual difference effects in younger learners when completing TOEFL Listen-Write tasks, the present participants seemed to be able to handle the multiple modes well and also did not report particular cognitive functioning difficulties affecting their task performance.

The factor analysis and correlation results pointed towards a single underlying construct in viewing-to-write tasks, rather than a sequence of separate modes/skills. The qualitative data, extensively demonstrating mode integrations (41% of codes), suggest this construct is likely to be multimodal integrated skills, and more specifically

viewing-to-write. While listening and writing skills are unmistakably involved in task performance, as evidenced by the interview data and medium effect sizes of the Aptis listening/writing predictor variables, they did not fully explain performance (product nor process). For example, the interviews revealed how visual and aural input interacted to establish meaning, or how visuals enabled content accuracy in writing – highlighting the tasks’ multimodal nature. Additionally, participants’ comments illustrated how transformation of the viewing input (versus just transcription of the aural input) is required to successfully complete the tasks, highlighting these tasks’ integrated nature. An example of transformation reported for the *viewing-to-describe* tasks was distinguishing main points from details in the video and connecting these into a coherent process description. For the *viewing-to-compare-and-contrast* tasks, an example was inferring what are similar/different/independent opinions in the video and signalling that understanding through cohesive devices in writing. In sum, our various datasets provide empirical evidence for our hypothesised operationalisation of a form of multimodal integratedness. Our research also shows that the tasks cannot be successfully completed through a sequence of separate modes and skills that operate distinctly, and that input processing cannot be separated from the production of output.

The observation that the Aptis listening scores correlated strongly with the viewing-to-write performance scores also confirms that integrated tasks should not be treated as primarily productive skills tasks, as has sometimes been done in the past. Sufficient input comprehension is needed, e.g., to use relevant lexis and effectively organise the information in writing (with respect to the targeted language function), just as sufficient writing skills are needed to demonstrate input comprehension through written discourse. Our finding seemingly conflicts with some prior studies on integrated tasks which found stronger score correlations with the productive rather than receptive independent skill(s) involved (see Plakans, 2022 for a review of research on this); however, such findings seem to result from tasks and rating scales with less strong operationalisations of integration. Our task instructions, however, require that the written performance exclusively represents selected content ideas from the video input, in relation to the targeted language function, rather than for example expressing an opinion on the input or generating new content ideas, as is required by some other integrated task types. Thus, our tasks demand that the written product demonstrates a direct connection with the listening-visual input, whereas in some other task types the writing component is not as closely related to the listening input which merely serves as content support or inspiration for writing. Additionally, as described in the Methodology section, our rating criteria systematically operationalise the multimodal integratedness – and thus also listening. Namely, the *Viewing for writing* criterion evaluates how well the content points from the video input are represented in the written product, and the *Organisation and Structure* and *Language Use* criteria make explicit connections back to the organisational and linguistics characteristics of the video input, rather than evaluating the linguistic characteristics of the writing independent of the input. Thus, listening (as part of the multimodal video input) is brought more strongly to the fore than in some other integrated scoring approaches where the productive skill is proportionally weighted more strongly. Furthermore, studies investigating integrated performances’ discourse characteristics and task completion processes have also evidenced the important role of receptive skills, and the use of transformation and discourse synthesis processes in integrated tests (see Plakans, 2022 for a review) – reflecting our findings.

As rating scales contribute to construct representation in performance assessments (Knoch et al., 2021), our rating scale design played a crucial role in representing multimodal integrated skills. While the finding of a single underlying construct might seemingly be in conflict with analytic rating, our scales' criteria and descriptors had multimodal transformation at their core throughout, e.g. *Viewing for Writing: Describe* task – accurate selection from and description of the different stages of the manufacturing process described in the video, *Language Use* – accuracy, range and complexity of grammar commensurate to that in the video input. The reference point for the content of the writing was thereby the video in its entirety of aural and visual modalities (including speech, pictures, gestures, body language, graphs, written words, etc.). Furthermore, advantages of analytic scales are that they prompt raters to systematically consider performances from various intended perspectives and offer more detailed information to learners/teachers – both useful in classroom L2 assessments (Kuiken & Vedder, 2021). However, given the distinct language functions targeted in *viewing-to-describe* and *viewing-to-compare-and-contrast* tasks, effective multimodal transformation and integration takes different shape. Thus, the *Viewing for Writing* and *Organisation and Structure* criteria were distinctly operationalised for the two task types, e.g. *Viewing for Writing: Describe* – accurate sequence of the manufacturing process stages represented in the video; *Compare-and-contrast* – fully correct attribution of views to each individual speaker.

Finally, we would like to share some practical task-design advice, to ensure assessing multimodal integratedness. First, as we know from video listening tests, the exact nature of the visuals and how they are included can determine whether or not input processing/comprehension goes beyond the aural channel only and thus operationalizes multimodality. Brett (1997), for example, reported half of test takers not looking up at the video. Our input design, however, seemed to have ensured visual processing as all participants reported watching the video (on at least one of the two plays). Our viewing inputs did not just display content pictures/drawings, but also written words/phrases (e.g. technical terms, key numbers), indicator shapes and moving animations to clarify materials and operations, and graphs/tables with additional data. The aural input also drew attention to the videos through phrases such as “let’s have a look at”, “in the graph here you can see”, “look how ...”, “just look at this picture!”. The *viewing-to-compare-and-contrast* inputs additionally showed the two speakers discussing, thus sharing body language (e.g. lip movements, facial/hand gestures like pointing to graphs or communicating emotions). Thus, we recommend using various visuals, making them slightly extend information conveyed through the auditory channel, and explicitly connecting visual and aural modes. Second, while single or double play of listening input is often debated, recent empirical research supports the validity of repeating input in single mode tasks (e.g. Holzknicht & Harding, 2023). We had hypothesised that, given increased complexity of multiple modes, double play would be justified in *viewing-to-describe* tasks too. Our data support this; the performance scores showed that the tasks are not easy, and during the recall interviews participants reported partial processing only after single play and needing double play for comprehensive processing, balancing attention and integrating all modes. Similarly, we anticipated that allowing notetaking is important to alleviate construct-irrelevant memory demands. Our recall data confirm this, and, crucially, also that notetaking operates as a means of integrating multimodal meaning. Thus, we advise double play of video input and allowing notetaking while watching. Last, we recommend giving test-takers a clear

direction for viewing, highlighting the target language function in the first instructions already, as the recall data showed that purposeful viewing (i.e. watching the video while knowing what will be expected at the writing stage) regulates and enhances integration of the multiple modes. We also advise presenting these instructions in both writing and speech in the video, to give test-takers the best opportunity to successfully understand what is expected of them.

CONCLUSION

This study was situated in the vastly underexplored area – in operational and scholarly terms – of multimodal integrated L2 formal assessment. Our findings, based on quantitative and qualitative data, indicate that our innovative viewing-to-write tasks elicit, reflect and evaluate this form of multimodal integrated language use and are practical for the formal assessment of such abilities of CEFR B-C level learners. Our work widens the repertoire of tasks and construct representation in L2 proficiency assessments, and demonstrates the feasibility of developing and administering multimodal integrated tasks in low-resource, classroom environments. Our freely available materials serve as guidance for formal assessment and development work in other settings. Our task specifications might also assist in writing effective prompts for recently emerging advanced AI systems such as large language models and graphics and moving image generation tools for creating video input for multimodal integrated viewing-to-write tasks. However, a human co-intelligence approach (Mollick, 2024) is required for carefully reviewing and revising the scripts and visuals generated by AI before they are used for assessment purposes. Without human oversight and co-creation, video input can represent inaccurate, biased or sensitive content (Goh & Aryadoust, 2025; Xi, 2023) and might underrepresent communicative and linguistic features of authentic language input (Sardinha, 2024).

As our study was restricted to one type of multimodal integrated task (viewing-to-write), we would encourage future work to experiment with alternative tasks. Also, while our operationalisation of multimodal integration was stronger than in most prior studies, this can still be increased. For example, while our input was multimodal and required integration into the output (a further mode), the latter on its own was limited to monomodal writing. Possibly, this could be extended to viewing-to-compose, with use of multiple modalities in the output too (e.g. visuals, videos), although there will be feasibility challenges for standardised testing. Additionally, in terms of rating, the present study comprised the development and application of rating scales for assessing viewing-to-write. While this work was thorough (involving extensive pre-piloting and piloting), a next logical step would be to conduct more sophisticated research on the rating scales' functioning, e.g. a rater cognition study.

DISCLOSURE STATEMENT

Tineke Brunfaut is co-editor of the British Council Monographs on Modern Language Testing. Judit Kormos is a member of the British Council's language teaching advisory committee.

FUNDING

The work was supported by the British Council through an Assessment Research Grant 2022. The British Council does not discount or endorse the methodology, results, implications or opinions presented by the researchers.

ORCID

Tineke Brunfaut  <http://orcid.org/0000-0001-8018-8004>

Judit Kormos  <http://orcid.org/0000-0002-2643-7222>

REFERENCES

- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beltrán-Palanquez, V. (2024). Assessing video game narratives: Implications for the assessment of multimodal literacy in ESP. *Assessing Writing*, 60, 100809. <https://doi.org/10.1016/j.asw.2024.100809>
- Brett, P. (1997). A comparative study of the effects of the use of multimedia on listening comprehension. *System*, 25(1), 39–53. [https://doi.org/10.1016/S0346-251X\(96\)00059-0](https://doi.org/10.1016/S0346-251X(96)00059-0)
- British Council. (2023). *Aptis guide for teachers November 2023*.
- Cheung, A. (2023). Developing and evaluating a set of process and product-oriented classroom assessment rubrics for assessing digital multimodal collaborative writing in L2 classes. *Assessing Writing*, 56, 100723. <https://doi.org/10.1016/j.asw.2023.100723>
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment - Companion volume*. Council of Europe Publishing. www.coe.int/lang-cefr
- Crawford Camiciottoli, B., & Campoy-Cubillo, M. C. (2018). Multimodal perspectives on English language teaching in higher education [Special issue]. *System*, 77, 1–9. <https://doi.org/10.1016/j.system.2018.03.005>
- Dunn, K. (2020). Aptis scoring system. *Aptis technical report*. TR/2020/002. https://www.britishcouncil.org/sites/default/files/aptis_scoring_system_v2.0.pdf
- Early, M., Kendrick, M., & Potts, D. (2015). Multimodality: Out from the margins of English language teaching [special issue]. *TESOL Quarterly*, 49(3), 447–460. <https://doi.org/10.1002/tesq.246>
- Goh, C. C. M., & Aryadoust, V. (2025). Developing and assessing second language listening and speaking: Does AI make it better? *Annual Review of Applied Linguistics*, 45, 179–199. <https://doi.org/10.1017/S0267190525100111>
- Hafner, C. A., & Ho, W. Y. J. (2020). Assessing digital multimodal composing in second language writing: Towards a process-based model. *Journal of Second Language Writing*, 47, 100710. <https://doi.org/10.1016/j.jslw.2020.100710>
- Holzknicht, F., & Harding, L. (2023). Repeating the listening text: Effects on listener performance, metacognitive strategy use, and anxiety. *TESOL Quarterly*, 58(1), 451–478. <https://doi.org/10.1002/tesq.3249>
- Knoch, U., Fairbairn, U., & Jin, Y. (2021). *Scoring second language spoken and written performance*. Equinox.
- Kormos, J., Brunfaut, T., & Michel, M. (2020). Motivational factors in computer-administered integrated skills tasks: A study of young learners. *Language Assessment Quarterly*, 17(1), 43–59. <https://doi.org/10.1080/15434303.2019.1664551>

- Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.
- Kuiken, F., & Vedder, I. (2021). Scoring approaches: Scales/rubrics. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 125–134). Routledge.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lim, F. V., Toh, M., & Nguyen, T. T. H. (2022). Multimodality in the English language classroom: A systematic review of literature. *Linguistics and Education*, 69, 101048. <https://doi.org/10.1016/j.linged.2022.101048>
- Liu, T., & Aryadoust, V. (2024). Does modality matter? A meta-analysis of the effect of video input in L2 listening assessment. *System*, 120, 103191. <https://doi.org/10.1016/j.system.2023.103191>
- Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learners' written performances. *Journal of Second Language Writing*, 45, 31–45. <https://doi.org/10.1016/j.jslw.2019.03.002>
- Mills, K. A. (2016). *Literacy theories for the digital age: Social, critical, multimodal, spatial, material and sensory lenses*. Multilingual Matters.
- Mills, K. A., & Unsworth, L. (2017). Multimodal literacy. *Oxford Research Encyclopedia of Education*. <https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-232>
- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Penguin Random House.
- Nelson, N., & King, J. R. (2022). Discourse synthesis: Textual transformations in writing from sources. *Reading and Writing*, 36(4), 769–808. <https://doi.org/10.1007/s11145-021-10243-5>
- O'Sullivan, B. (2015). Linking the Aptis reporting scales to the CEFR. *Technical report tr/2015/003*. British Council.
- O'Sullivan, B., Dunlea, J., Spiiby, R., Westbrook, C., & Dunn, K. (2020). Aptis general technical manual, version 2.2. *Technical report tr/2020/001*. British Council.
- Palmour, L. (2024). Assessing speaking through multimodal oral presentations: The case of construct underrepresentation in EAP contexts. *Language Testing*, 41(1), 9–34. <https://doi.org/10.1177/02655322231183077>
- Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.01109>
- Peters, E., & Muñoz, C. (2020). Language learning from multimodal input. *Studies in Second Language Acquisition*, 42(S3), 489–665.
- Plakans, L. (2022). Writing integrated tasks. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed. pp. 357–371). Routledge.
- Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, 102(4), 713–731. <https://doi.org/10.1111/modl.12509>
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing (4.0.2).
- Rossi, O., & Brunfaut, T. (2021). Text authenticity in listening assessment: Can item writers be trained to produce authentic-sounding texts? *Language Assessment Quarterly*, 18(4), 398–418. <https://doi.org/10.1080/15434303.2021.1895162>
- Sardinha, T. B. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1), 100083. <https://doi.org/10.1016/j.acorp.2023.100083>
- Shohamy, E. (2022). Critical language testing, multilingualism and social justice. *TESOL Quarterly*, 56(4), 1445–1457. <https://doi.org/10.1002/tesq.3185>

- So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). Toefl Junior® design framework. *ETS Research Report Series*, 2015, 1–45. <https://doi.org/10.1002/ets2.12058>
- Tan, L., Zammit, K., D'warte, J., & Gearsides, A. (2020). Dialogic inquiry in multimodal literacy education: Joining the dots between assessment and literacy curriculum [Special issue]. *Language and Education*, 34(2), 97–114. <https://doi.org/10.1080/09500782.2019.1708926>
- Weir, C. J. (1990). *Communicative language testing*. Prentice Hall.
- Whithaus, C. (2014). Multimodal assessment [Special Issue]. *Computers and Composition*, 31, 1–86.
- Xi, X. (2023). Advancing language assessment with AI and ML-learning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, 20(4–5), 357–376. <https://doi.org/10.1080/15434303.2023.2291488>
- Yi, Y., Shin, D., & Cimasko, T. (2020). Multimodal composing in multilingual learning and teaching contexts [Special issue]. *Journal of Second Language Writing*, 47, 100717. <https://doi.org/10.1016/j.jslw.2020.100717>
- Yu, G., Rea-Dickins, P., & Kiely, R. (2010). The cognitive processes of taking IELTS academic writing task 1. *IELTS Research Reports* (Vol. 11). IELTS.
- Zhang, M., Akoto, M., & Li, M. (2023). Digital multimodal composing in post-secondary L2 settings: A review of the empirical landscape. *Computer Assisted Language Learning*, 36(4), 694–721. <https://doi.org/10.1080/09588221.2021.1942068>

Appendices

Appendix A

Viewing-to-write rating scales (downloadable from https://osf.io/b3jr5/?view_only=7793563fb6f048439cd8a037acaca4dd)

RATING SCALE – Describe a process

	Viewing for writing* <ul style="list-style-type: none"> Selection of relevant process steps from the video Accurate description of process steps from the video Accurate sequence of process steps from the video 	Organisation & structure <ul style="list-style-type: none"> Text organisation (beginning-mid-ending) Flow of process description Use of signposting, including cohesive devices to reflect process 	Language use <ul style="list-style-type: none"> Accuracy, range and complexity of grammar Accuracy, range and sophistication of vocab Effect of linguistic errors on understanding 	Mechanics <ul style="list-style-type: none"> Spelling Control of punctuation
4	- All relevant process steps selected from the video - Fully accurate description of the process steps - Fully accurate sequence of the process steps E.g. 8/8 content points	- Fully effective organisation of the text, including a highly effective beginning-mid-ending - Fully effective flow of process description - Fully effective signposting, including excellent use of a solid range of cohesive devices to indicate process steps and sequence	- Grammar: very high level of accuracy; wide range and high complexity commensurate to the video input - Vocab: very high level of accuracy; wide range and high sophistication commensurate to the video input - Linguistic errors: if any, they do not impede understanding	2 - Flawless spelling - Excellent control of punctuation
3	- 1 relevant process step missing - Inaccurate description of 1 process step, or partly inaccurate description of up to 2 process steps - Accurate sequence of the process steps represented E.g. 7/8 content points	- Good organisation of the text, including an appropriate beginning-mid-ending - Good flow of the process description - Good signposting, including appropriate use of a number of different cohesive devices to indicate process steps and sequence	- Grammar: good level of accuracy; appropriate range and complexity commensurate to the video input - Vocab: good level of accuracy; appropriate range and sophistication commensurate to the video input - Linguistic errors: if any, they hardly impede understanding	1.5 - Good spelling (very few errors) - Good control of punctuation (very few errors)
2	- 2 or 3 relevant process steps missing - Inaccurate description of 2 or 3 process steps, or partly inaccurate description of 3 or 4 process steps, or a combination of these - Some inaccurate sequencing of the process steps represented E.g. 6/8 or 5/8 content points	- Acceptable textual organisation; beginning or ending might be missing or rather abrupt - Partly jumpy process description - Acceptable signposting overall; use of a more limited range of cohesive devices, the odd one used incorrectly, to indicate any process steps and sequence	- Grammar: acceptable level of accuracy (may include systematic errors); acceptable range and complexity commensurate to the video input - Vocab: acceptable level of accuracy (may include systematic errors); acceptable range and sophistication commensurate to the video input (may contain frequent repetition) - Linguistic errors: sometimes impede understanding	1 - Acceptable spelling (occasional errors) - Reasonable control of punctuation (occasional errors or missing punctuation)
1	- Half to ¼ of relevant process steps missing - Inaccurate description of half to ¼ of process steps, or partly inaccurate description of 5 or more process steps, or a combination of these - Inaccurate sequencing of several of the process steps represented E.g. 4/8, 3/8 or 2/8 content points	- Limited or poor organisation of the text; no beginning-mid-ending structure - Jumpy process description - Limited or poor signposting; use of high-frequency cohesive devices only, which are repeated several times and/or used incorrectly, to indicate any process steps and sequence	- Grammar: limited level of accuracy; restricted range and complexity (also relative to the video input) - Vocab: limited level of accuracy; restricted range and sophistication (also relative to the video input) - Linguistic errors: frequently impede understanding	0.5 - Weak spelling (frequent errors) - Limited control of punctuation (frequent errors or missing punctuation)
0	- Only 1 or no relevant process steps selected from the video - Any process steps included inaccurately described - Any process steps included inaccurately sequenced - No performance E.g. 1/8 or 0/8 content points	- Lack of textual organisation - Lack of process description - Lack of signposting and cohesive devices to indicate any process steps and sequence - No performance	- Grammar: inadequate evidence of accuracy, range and complexity - Vocab: inadequate evidence of accuracy, range and sophistication - Linguistic errors: seriously impede any understanding - No performance	0 - Poor spelling throughout - Poor punctuation throughout - No performance

*See crib sheet for the process steps of each prompt.

NOTE: A performance might show features from different score bands (some from a higher and some from a lower band). Please make up the balance to award the higher or lower band.

RATING SCALE – Compare and contrast speakers' views

	Viewing for writing* <ul style="list-style-type: none"> Comprehensive representation of views Accurate representation of speakers' views Correct attribution of views to individual speaker 	Organisation & structure <ul style="list-style-type: none"> Text organisation (beginning-main body) Flow of comparing and contrasting Use of signposting, including cohesive devices and referents to compare and contrast speakers' views 	Language use <ul style="list-style-type: none"> Accuracy, range and complexity of grammar Accuracy, range and sophistication of vocab Effect of linguistic errors on understanding 	Mechanics <ul style="list-style-type: none"> Spelling Control of punctuation
4	- All speaker views from the video represented - Fully accurate representation of speakers' views - Fully correct attribution of views to each individual speaker E.g. 4/4 content points	- Fully effective organisation of the text, including a highly effective beginning and main body - Very well-flowing comparing and contrasting - Fully effective signposting, including excellent use of a solid range of cohesive devices and of explicit referents to compare and contrast speakers' views	- Grammar: very high level of accuracy; wide range and high complexity commensurate to the video input - Vocab: very high level of accuracy; wide range and high sophistication commensurate to the video input - Linguistic errors: if any, they do not impede understanding	2 - Flawless spelling - Excellent control of punctuation
3	- 1 content point with one or both speaker views missing - Inaccurate representation of 1 content point in terms of one or both speaker views - Incorrect attribution in 1 content point of one or both views to each individual speaker E.g. 3/4 content points	- Good organisation of the text, including an appropriate beginning and main body - Good flow of comparing and contrasting - Good signposting, including appropriate use of a number of different cohesive devices and of referents to compare and contrast speakers' views	- Grammar: good level of accuracy; appropriate range and complexity commensurate to the video input - Vocab: good level of accuracy; appropriate range and sophistication commensurate to the video input - Linguistic errors: if any, they hardly impede understanding	1.5 - Good spelling (very few errors) - Good control of punctuation (very few errors)
2	- 2 content points with one or both speaker views missing - Inaccurate representation of 2 content points in terms of one or both speaker views - Inaccurate attribution in 2 content points of one or both views to each individual speaker E.g. 2/4 content points	- Acceptable textual organisation; beginning might be missing or rather abrupt - Partly jumpy comparing and contrasting - Acceptable signposting overall; use of a more limited range of cohesive devices to compare and contrast speakers' views, with perhaps the odd cohesive device used incorrectly; 1 or 2 ambiguous or missing referents to signal an individual speaker	- Grammar: acceptable level of accuracy (may include systematic errors); acceptable range and complexity commensurate to the video input - Vocab: acceptable level of accuracy (may include systematic errors); acceptable range and sophistication commensurate to the video input (may contain frequent repetition) - Linguistic errors: sometimes impede understanding	1 - Acceptable spelling (occasional errors) - Reasonable control of punctuation (occasional errors or missing punctuation)
1	- 3 content points with one or both speaker views missing - Inaccurate representation of 3 content points in terms of one or both speaker views - Inaccurate representation in 3 content points of one or both views to individual speakers E.g. 1/4 content points	- Limited or poor organisation of the text; no beginning-main body structure - Jumpy comparing and contrasting - Limited or poor signposting; use of high-frequency cohesive devices only, which are repeated several times and/or used incorrectly, to compare and contrast speakers' views; several ambiguous or missing referents to signal an individual speaker	- Grammar: limited level of accuracy; restricted range and complexity (also relative to the video input) - Vocab: limited level of accuracy; restricted range and sophistication (also relative to the video input) - Linguistic errors: frequently impede understanding	0.5 - Weak spelling (frequent errors) - Limited control of punctuation (frequent errors or missing punctuation)
0	- Only 1 or no views from the video represented - Any views included inaccurately represented - Any views included inaccurately attributed to individual speakers, or lack of attributions E.g. 0/4 content points	- Lack of textual organisation - Lack of comparing and contrasting - Lack of signposting, cohesive devices and referents to compare and contrast speakers' views - No performance	- Grammar: inadequate evidence of accuracy, range and complexity - Vocab: inadequate evidence of accuracy, range and sophistication - Linguistic errors: impede all understanding - No performance	0 - Poor spelling throughout - Poor punctuation throughout - No performance

*See crib sheet for the speaker views in each prompt.

NOTE: A performance might show features from different score bands (some from a higher and some from a lower band). Please make up the balance to award the higher or lower band.

Appendix B

Inter-rater reliability (Spearman's rho; $n_{\text{performances}} = 30$)

	Whole scale	Viewing for writing	Organisation & structure	Language use	Mechanics
Describe a process scale	.94**	.90**	.88**	.79**	.65**
Compare-and-contrast scale	.77**	.79**	.67**	.75**	.76**

** Significant at the 0.01 level.

Appendix C

Descriptive statistics for the viewing-to-write tasks

	Min	Max	Mean	SD	Skewness	Kurtosis
<i>All four tasks ($n_{\text{participants}}=134$; $n_{\text{performances}}=357$)</i>						
Viewing for writing	.00	4.00	2.62	.94	-.715	.183
Organization & structure	.50	4.00	2.73	.78	-.885	1.180
Language use	1.00	4.00	2.68	.68	-.582	.636
Mechanics	.25	2.00	1.21	.36	-.386	-.363
Writing composite total	1.00	10.00	6.62	1.58	-.683	.735
Total score	1.00	13.75	9.23	2.36	-.857	.863
<i>Describe task: Contact Lenses ($n=93$)</i>						
Viewing for writing	.00	4.00	2.62	.99	-.407	-.569
Organization & structure	.50	4.00	2.74	.70	-1.162	2.239
Language use	.00	4.00	2.64	.69	-.944	1.520
Mechanics	.25	1.75	1.16	.37	-.355	-.577
Writing composite total	1.50	9.25	6.53	1.52	-.836	1.020
Total score	2.00	13.00	9.15	2.32	-.825	.713
<i>Describe task: Train ($n=88$)</i>						
Viewing for writing	.00	4.00	2.65	.86	-1.179	1.780
Organization & structure	.00	4.00	2.72	.80	-1.087	1.550
Language use	.50	4.00	2.48	.73	-.304	.244
Mechanics	.25	2.00	1.17	.36	-.140	-.234
Writing composite total	1.00	10.00	6.39	1.65	-.646	.892
Total score	1.00	13.50	9.04	2.35	-1.075	1.756
<i>Compare & Contrast task: Zoo ($n=91$)</i>						
Viewing for writing	.00	4.00	2.67	1.00	-.953	.576
Organization & structure	.00	4.00	2.75	.75	-.933	1.741
Language use	1.00	4.00	2.86	.62	-.435	.354
Mechanics	.25	2.00	1.32	.35	-.840	.983
Writing composite total	2.25	9.75	6.93	1.48	-.748	1.007
Total score	2.25	13.50	9.60	2.33	-1.007	1.199
<i>Compare & Contrast task: Space Tourism ($n=85$)</i>						
Viewing for writing	.00	4.00	2.50	.95	-.478	-.181
Organization & structure	.50	4.00	2.69	.86	-.492	.063
Language use	1.00	4.00	2.74	.63	-.452	.290
Mechanics	.25	1.75	1.18	.37	-.280	-.745
Writing composite total	2.00	9.75	6.61	1.64	-.525	.327
Total score	2.50	13.75	9.14	2.46	-.622	.324

Appendix D

Descriptive statistics for the Aptis Test ($n = 134$)

	Min	Max	Mean	SD	Skewness	Kurtosis
Listening	18	50	42.30	4.86	-1.613	4.776
Writing	28	48	44.13	3.56	-2.131	6.147
Vocabulary & Grammar	17	49	39.17	5.65	-1.180	1.873

Appendix E

Frequency of processing codes for the viewing-to-write performances

	All four tasks ($n=20$)	Describe: Contact Lenses ($n=10$)	Describe: Train($n=10$)	Compare & Contrast: Zoo ($n=10$)	Compare & Contrast: Space Tourism ($n=10$)
Watching visuals	115 (10%)	22 (9%)	34 (11%)	33 (12%)	26 (9%)
Listening	336 (30%)	83 (35%)	69 (22%)	85 (31%)	99 (35%)
Writing	195 (18%)	41 (17%)	59 (20%)	50 (19%)	45 (16%)
Integration: audio & visual	60 (5%)	12 (5%)	26 (8%)	11 (4%)	11 (4%)
Integration: input to notes	233 (21%)	39 (16%)	56 (18%)	67 (25%)	71 (25%)
Integration: input to writing	91 (8%)	24 (10%)	41 (13%)	11 (4%)	15 (5%)
Integration: notes to writing	73 (7%)	19 (8%)	26 (8%)	13 (5%)	15 (5%)
<i>Total</i>	<i>1103 (100%)</i>	<i>240 (100%)</i>	<i>311 (100%)</i>	<i>270 (100%)</i>	<i>282 (100%)</i>